

博士論文

**Search and Communication Based on
Affective Understanding of Fonts and Images**

(画像とフォントの印象理解に基づく
検索・コミュニケーション支援)

崔 セミ

Acknowledgements

First, and foremost, I would like to express my sincere gratitude to my advisor Professor Aizawa for the continuous support during my Ph.D. study at the University of Tokyo. Without your support and supervision, I have not been able to start and continue this study for last 7 years. You gave me a lot of opportunities to grow up. I truly respect you not only as a researcher but also as my guide through life. The wisdom, passion, openness, and kindness you showed let me know how a researcher who has one's mission serves academic as well as society. I also really enjoyed having small-talk with you.

I also give big thanks to Professor Yamasaki. I was really impressed by your keen comments at every meeting and those comments made my research better and myself growing so much. You encouraged me to attend conferences to see the bigger world and motivated me a great deal. There were so many things I learned from you, from constant support for your students to your tidy and diligent life. I always remember you as a great professor throughout my master and Ph.D. course.

I also appreciate Matsubayashi-san for everything including lab works as well as consultations. Without your help, nothing could have gone smoothly. I will never forget your help from research related works such as attending a conference and preparing exchange student to lots of supportings for living as an international student in Japan. Your dedications to your family and our lab members were respectful.

I would like to take this opportunity to acknowledge all the members in Aizawa-Yamasaki laboratory. As I started my life as a research student in 2012 April, the first lab members I met were Ikehata-san and Matsui-san. I remember the day. It was a little early morning and the lab was empty, but there were only two you all (or a few more). And now, you all still have the same, or the bigger passion to research. I guess I am so lucky to have senpai like you all. All the best and I hope to see you all in future at some random academic venue too. To Onkar, I am so happy to get to know you and spent the Ph.D. time together. I could get through all the time with your positive energy and kind heart. It was also good to do the collaboration project together. I was just happy being able to do something together as a douki. I also thank the multimedia and manga group members; Mari-chang, Matsumura-kun, Ikuta-kun, Narita-kun, and Tsubota-kun. Though it was not that long time to have a meeting together, the discussions and sparkling ideas you all showed really motivated me a lot. To Furuta-kun and Ikami-kun, you two are probably the oldest lab members I spent time together except for the professors. I am very lucky to have such a talented kohai. There is no doubt that you all will be

a great researcher to lead Japan. I hope all the best in your future workplace. Especially Furuta-kun, thank you very much for being kind to me since all my douki left. The excursion to the sea with Midorikawa-kun and Ito-kun is one of my precious memory in Japan. Sincerely. To my lovely girls, Mari, Hu and Yiwei, I was so happy to be with you all. The time for sweets, meals and every talk we had made me mentally relieved. To Hu, my first swimming class student, I am so proud that you can swim in the Beijing flood. Mari-chang, I will not forget your warm heart and mild personality. Yiwei, there will not be a cute girl like you in this world. I feel like I received a lot from you all. Hope there will be some chance to repay your kindness.

I cannot help appreciating to all the MHUG members; Gloria, Enver, Radu, Synziana, Mercedes, Ciara, Wei, Dan, Aliaksandr. I really miss you guys. The time I spent in Trento with you all refreshed me, and it is so precious to me. Gloria, I really admire the way you see the world. Enver, you made me realize the beauty of Italy. Radu and Synziana, I was impressed by your love for nature. Since then, I feel like I have larger love for my furry friend. Mercedes, your positive energy and warm heart encouraged me a lot. Wei, I was so happy to meet the author of my favorite paper. Lastly, Nicu, I was able to set up a new project supported by you. You gave me a chance to get valuable comments and insight into the project. I am looking forward to seeing you again.

I would like to take this opportunity to appreciate and dedicate this dissertation to my family. Since I decided to study in Japan, you all followed my decision and supported me a lot. Whenever I visit my home, you made me special food and time. When I am stressed out, you kept sending me lots of pictures of my lovely dog Hodu. But above all, just knowing your presence gives me energy. I am so happy for our reunion after my graduation.

To my best friends from Todai, Jeonghyun, Hyemin, Sojin, Gwangnam, Soomin, and our baby angel Sarang-chang. I will not forget the summer's days we ran together to catch Pokemon in Odaiba. I am not the Pokemon trainer anymore, I know that the time for lunch, home parties, soju in Jingogae will continue in the future. My Ph.D. life is filled with time with you. Especially to Jeonghyun, I know how much we are different from each other. But the way you see the world gave me another perspective and inspired me to see things differently. You are like my muse. I also know how similar our tastes we have. It was my daily pleasure to find the right food for us every single day. Every small talk with you was full of joy. But most of all, I will not forget the times we prayed for each other. I am so happy to have you as my life long rival, friend, and family.

Lastly, I thank my Load for being with me from the beginning to the end of the study. I can not count how many blessings I have got from you. Every achievement during my course is done by you not me.

Saemi Choi

Feb 4, 2019

Abstract

Graphic elements, such as images and fonts, are good visual communication media. These media convey a variety of information, and they affect one's emotional state. A photo of sunset creates a feeling of warmth, and a text with the font Times New Roman would look professional. With the growth of World Wide Web, significant amount of images and fonts have been shared via on-line communities. Many studies have proposed applications that support users to search images and fonts in the large database, but few studies considered affect.

Motivated by these observations, we aim at modeling systems that understand affective signals in image and font and proposing applications where the analyzed affective signals can be used. In this thesis, we explore the answers to the following research questions: (1) How to model a system that predicts affects in images and fonts without a large dataset to learn (2) How to improve user engagement in searching with ambiguous and noisy dataset? This thesis explores answers to these questions as follow:

Image impression retrieval: Conventional image retrieval systems ask users to input query by text. However, it is not always easy for users to convert their intention into verbal representations. In Chapter 3, we propose an interactive retrieval system based on yes-no questions for image impression retrieval. We modeled a system that interprets images with impression words such as fresh and modern. Then, we introduce a yes-no question based querying method and a feedback interface to support users querying.

Font emotion understanding: Different fonts create different experience. Many researchers in marketing field studied the effect of fonts in advertisement, but, few researchers studied emotional effects of font. In Chapter 4, we demonstrate the effect of fonts on viewers' emotional state by two experimental studies — explicit study and implicit study. In explicit study, we measure the response to fonts using a questionnaire method. In implicit study, we measure unconscious responses to fonts by analyzing spontaneous speeches that elicited by different fonts.

Font communication on mobile messenger: Instant messaging is a popular form of text-based communication. However, text-based messaging lacks the ability to communicate nonverbal signals such as facial expressions and tones of the voice. In Chapter 5, we propose Emotype, a mobile messenger application prototype that enables users to change the font of their message to

communicate emotions. In user test, we demonstrate the feasibility of fonts for communicating emotions, and understand user experience with the application.

Font search by image: One of the important aspects in graphic design is choosing the font of the caption that matches aesthetically the associated image. In Chapter 6, we present two font search systems that enable users to use images as queries - (1) query by image impressions based on color study and (2) query by image contents based on concept analysis. Instead of matching font and image directly, we mapped both image and font to color-based semantic space or concept-based semantic space. Our evaluation results show that the recommended fonts scored better than other comparisons and provides competing results with the ones chosen by experienced graphic designers.

Creativity support in graphic design: Inspiration plays an important role in the creative process. By getting inspired, we can reach unexpected but useful ideas. Inspiration, generally, comes to us when we interact with external interventions. In Chapter 7, we present a framework that assist users' interactions in font search with unexpected but useful concepts generated by multimodal learning. By examining the results of the model that change with various inputs, we observed that the model produces promising results that appeared to be useful for inspiring users.

In this thesis, we aim at modeling systems that understand affective signals in image and font and proposing applications where the analyzed information can be used. Especially, we see machines as a medium for affective interaction between users, and focus on studying the interactional influence of the system to users to make users be pleased in the affects-aware systems.

Table of contents

List of figures	x
List of tables	xiv
1 Introduction	1
1.1 Affective Signals in Image and Font	2
1.2 Advances in Affective Computing	3
1.2.1 Two perspectives on affective computing: Cognitive vs Interactional approach	3
1.2.2 Scale of affects	4
1.3 Research Challenges	5
1.4 Objective and Contributions	5
2 Related Works: Research and Development on Font over the Centuries	7
2.1 History of Font	8
2.2 Effect of Font	9
2.2.1 Readability of written text	9
2.2.2 Personality of written text	10
2.3 Font in Computing	11
2.3.1 Visual font recognition	11
2.3.2 Font search	11
2.3.3 Font generation	13
2.3.4 Font as tones of voice	14
3 Image Impression Learning for Retrieval	16
3.1 Introduction	17
3.2 Related Works	18
3.3 System Overview	19
3.4 Image Impression Analysis	19
3.4.1 Kobayashi color image scale	21
3.4.2 Impression word scale	21
3.4.3 Model training	21

3.4.4	Impression score	23
3.5	Querying Support	23
3.5.1	Yes-no questions	24
3.5.2	Feedback slide	26
3.6	Experiment	27
3.6.1	Task design and dataset	27
3.6.2	Task design	28
3.6.3	Results	29
3.7	Discussion	29
3.7.1	Proposed vs Text-based	30
3.7.2	Semantic agreement	30
3.8	Conclusions	33
4	Font Emotion Understanding: Measuring Explicit and Implicit Emotional Responses to Font	34
4.1	Introduction	35
4.2	Related Works	36
4.2.1	Font and emotion	36
4.2.2	Psychological emotion modeling theories	36
4.2.3	Font as tones of voice	37
4.3	Font Dataset	38
4.4	Explicit Testing	39
4.4.1	Experiment design	39
4.4.2	Results	41
4.4.3	Discussion	41
4.5	Implicit Testing	44
4.5.1	Experiment design	44
4.5.2	Paragraph level analysis	46
4.5.3	Word level analysis	49
4.5.4	Subjective evaluation on the collected speeches	50
4.6	Conclusions	52
5	Emotype: Expressing Emotions by Changing Typeface in Mobile Messenger Texting	53
5.1	Introduction	54
5.2	Related Works	55
5.2.1	Nonverbal signals in mobile messengers	55
5.2.2	The usage of typeface in multimedia	56
5.2.3	Typography communication in mobile messengers	57
5.2.4	Studies on affect and emotion	57

5.3	Material	59
5.3.1	Emotional typefaces	59
5.3.2	Messenger application prototype	60
5.4	Quantitative Study: Feasibility of Using Typefaces for Emotion Communication . . .	60
5.4.1	Task	61
5.4.2	Results	63
5.5	Qualitative Study: Exploring User Experiences	64
5.5.1	Method	64
5.5.2	Role-playing study	65
5.5.3	Focus group discussion	67
5.6	Discussion and Conclusions	70
6	Font Search by Image based on Color-based and Concept-based Matching Methods	72
6.1	Introduction	73
6.2	Related Works	74
6.2.1	Font search	74
6.2.2	Word embedding in search	74
6.3	Affective Color-based Matching	75
6.3.1	Methodologies	75
6.3.2	Experiment	81
6.3.3	Results and discussion	81
6.4	Affective Concept-based Matching	83
6.4.1	Methodologies	83
6.4.2	Experiment	92
6.4.3	Results and discussion	94
6.5	Limitations	95
6.6	Conclusions	95
7	Assist Users' Interactions in Font Search with Unexpected but Useful Concepts Generated by Multimodal Learning	97
7.1	Introduction	98
7.2	Related Works	100
7.2.1	Machines and creativity	100
7.2.2	Affective effects of font	101
7.2.3	Multimodal deep boltzmann machine	102
7.2.4	Inspiration	102
7.3	Methodologies	102
7.3.1	Dataset	103
7.3.2	Multimodal deep boltzmann machine	103

7.3.3	Font visual features	106
7.3.4	Model architecture for font image and tag MDBM	109
7.4	Qualitative Analysis	109
7.4.1	Tag to tag	109
7.4.2	Image to image	109
7.4.3	Image to tag	110
7.4.4	Tag to image	110
7.4.5	Tag and image to tag and image	112
7.4.6	User test	113
7.5	Quantitative Experiment	114
7.5.1	Evaluation	115
7.5.2	Results	115
7.6	Conclusions	116
8	Conclusion	117
8.1	Summary	118
8.2	Future Directions	119
	References	120

List of figures

1.1	Two images that are the same in content (house, sky, grassland), but feel different. . .	2
1.2	Two memos that are the same in text, but feel different.	3
2.1	Font development over the centuries	8
2.2	Visual font recognition	10
2.3	Font search with clear user preference	11
2.4	Font search without clear user preference	12
2.5	Font shape learning by decomposing the shape of characters	12
2.6	Semi-automatic font generation that allow user intervention. Users need to adjust skeleton of the input glyph and select semantic parts [142]	13
2.7	Font styling	13
2.8	Applications that see font as a visual medium that conveys tones of voice	14
3.1	An image pair (b) and (c) which have similar contents but different impressions is difficult to distinguish while (a) and (b) is easier to distinguish.	17
3.2	Proposed image retrieval system	19
3.3	Model training	20
3.4	Kobayashi color image scale on the Semantic Space	20
3.5	Processing of fuzzy answer from user	25
3.6	Processing of fuzzy answer from user	25
3.7	Feedback slide	27
3.8	Example images for the <i>Sweet and Dreamy Wedding</i> task (solid line images were collected by suggest system and broken line images were text-based system). Majority of positive images were collected by the proposed system.	29
3.9	Image Distribution on the Semantic Space (x-axis: Warm-Cool, y-axis: Soft-Hard). (1) shows the distribution of all images tagged with a task noun, e.g., <i>wedding</i> . (2) shows the distribution of images selected by the text-based system. (3) shows the distribution of images selected by the proposed system. The distribution of (2) is similar to (1) but (3) is significantly different from (1) or (2). It implies that the images retrieved by the text-based system are not well differentiated from that of total images.	31

3.10	Image Distribution on the Semantic Space (Proposed system), x-axis: warm-cool, y-axis: soft-hard. The average Euclidean distance is shortest with our system (blue colored) in most cases, and it implies that the images collected by proposed system has best consensus with Kobayashi model.	32
4.1	Examples of a text varying fonts. Texts can appear to convey negative feelings or positive feelings depending on the font used.	35
4.2	Russell's circumflex model	37
4.3	Example user assessment sheet. There are three questions for each font. Q1 and Q2 are for the dimensional emotion space (valence-arousal), and Q3 is for the categorical emotion state (six basic emotions).	39
4.4	The overall tendency of user assessments. The color of each point indicate the user-labeled emotion category (Angry(red), Disgust(green), Fear(purple), Happiness(yellow), Sadness(blue), Surprise(Black), Neutral(grey)).	40
4.5	Heat map for each emotion to highlight the region where the dimensional assessment was concentrated.	41
4.6	The example fonts with low agreement.	42
4.7	High consensus font examples for each emotional categories	43
4.8	Three fonts that we used in our implicit testing.	45
4.9	An example of a positive story and two analysis method (paragraph level and word level).	47
4.10	An example of the assessment sheet	50
5.1	User interface of the Emotype prototype. A user can send a text message and change the message typeface to convey a particular feeling.	54
5.2	Example of a two-alternative forced choice (2AFC) question.	58
5.3	Example of a two-alternative forced choice (2AFC) question.	58
5.4	Representative typefaces for each emotion category.	59
5.5	An envisioned interface for changing the typeface.	60
5.6	Examples of conversations for testing each hypothesis.	61
5.7	Example conversation task (3. <i>Winning the Lotto</i>)	64
5.8	How well the recipients B guessed the sender A's emotion depending on the typeface used. The sum of each row is one ($p < .001$ by Fisher's exact test). As we can see the highlighted results (red-dotted boxes), users reported positive feelings with both the neutral typeface and positive typeface with high probability.	66

5.9	The actual user responses in the vignettes 1. <i>Magazine Subscription</i> . We observe that the use of the positive typeface elicited an active response in row (d). We highlighted emphasizing expressions in bold, and colored the positive words green. For example, in row P1, with the neutral typeface, P1 simply gave a positive response (<i>Nice,</i>), and asked (<i>how do they like it?</i>). However, with the positive typeface, P1 showed a more active positive response (<i>So nice;</i>) and affirming (<i>they must like it</i>). Here, P_i indicates the id of each participant ($i=1,2,3,\dots, 8$).	68
5.10	An example using three different types of emotional expression: Emoji, Emotype and emotion word.	68
5.11	Messages showing inconsistency between the content and the typeface.	69
5.12	An free chat example which shows dynamic tone changes.	70
6.1	The flower image with soft color tones conveys soft and delicate feelings. We can say that the thin serif font on the left matches well the image as it gives more similar feelings than the thick and angular font on the right.	73
6.2	Proposed two font search methods that use two different semantic spaces.	74
6.3	Overview of the system flow. If a user inputs an image, a image description is generated by object recognition [Google], impression analysis [32] and natural language generating models [51] in panel 1. Then our proposed font impression model recommends fonts which have similar feelings to the given image (panel 2). Based on the recommendation, the system generates a description for supporting the recommendation (panel 3). Users can give feedback by controlling font features such as <i>slanted</i> (panel 4), and the results are updated in panel 5.	77
6.4	The font mapping results on the semantic space. One blue dot indicates one font. The five images were plotted in the semantic space by the image impression model. The phrase “Handgloves” is a filler text to demonstrate the shape of fonts.	79
6.5	Examples of font-image pairing task results	80
6.6	System overview. Users can search font using text query as well as image query. . .	83
6.7	Overview of font dataset. We finally collected total 8340 fonts which has 2,340 size of font tag dictionary.	84
6.8	Query images and top predicted concepts	85
6.9	Five movie theme posters (a) <i>How to Lose a Guy in 10 Days</i> , (b) <i>Soldiers of Fortune</i> , (c) <i>Annabelle</i> , (d) <i>Alice in Wonderland</i> , (e) <i>Earth to Echo</i>	93
6.10	An example task	93
7.1	Example of font image and associated tags. It consists of heterogeneous subjective information — the impression the font creates (gray), its application (yellow), and the typographical feature of the font (green).	98

7.2	The proposed framework is divided into three parts: building the dataset, font-feature learning, and multimodal learning. We collected large sets of a font with the associated tags and then used them to learn the visual features of the fonts and associated text.	99
7.3	A multimodal DBM (left) and two modality-specific DBMs (right two). The text-specific and image-specific DBM are used for modeling sparse word count vectors and real-valued dense image features respectively. There is one additional layer (green) that combines the two modality-specific models on top of them.	101
7.4	Five path variations across the model. Given the text input (tag), the model can reconstruct the tag conditioned on the input (a) Similarly, if we input the image, the model retrieves images that share a visual characteristic that is similar to the given input image. On passing through the joint layer, the model outputs a different modality from the input (c, d, and e).	103
7.5	Example of retrieved images of a given single topic (diagonal) and two tags from different topics (lower triangular) on following the pathway illustrated in Figure 7.4c.	110
7.6	Expanded tags from anchored tag and varying font image input following the pathway illustrated in Figure 4(e).	113

List of tables

3.1	30 Color Features (color values that defined by two space, CIELab and HSV)	22
3.2	5 Color Features that calculated using semantic space coordinations	23
3.3	Experimental results showing the average number of positive-group images for each query. In most cases (nine-twelfth), the number of positive-group images collected by the proposed system is more than text-based system. It indicates that the proposed system was more satisfactory for the users.	28
4.1	The label and the number of collected fonts for each category.	38
4.2	The user assessments results of high consensus font examples (Figure 4.7) and the user assessments results of low consensus font examples (Figure 4.6). Ha, Sa, Su, Fe, An and Di are abbreviation for Happiness, Sadness, Surprise, Fear, Angry and Disgust respectively. Here, we exclude the number of votes for neutral.	44
4.3	An example of negative story. it is composed of three paragraphs, i.e., background, focus and target. After giving reading material, we provided comprehension test to make participants to engage in reading task.	45
4.4	The result of our prosodic feature analysis. The higher score indicates that a certain prosodic feature (REC, RDC, RPC, RIC, or WPC) was more pronounced in the modulated by Font stimuli or Instructed stimuli than that of Normal. We can see that implicit responses by font stimuli seem to follow the tendency of explicit responses by instructed stimuli (* : significance@0.05 level in Binomial test).	48
4.5	The subjective evaluation from the crowd-sourced study. The number in brackets refers to the number of assessments. (* : significant@0.05 level in Binomial test) . .	51
5.1	The conversations provided for testing the two hypotheses. The lines in bold for each conversation indicate the target messages that have variations in the typeface used. For $H0_1$, we designed the content of the target message to be neutral with three typeface variations (neutral, positive, and negative). For $H0_2$, the target message was obviously negative or positive, and each has two typeface variations (neutral and negative vs. neutral and positive).	62

5.2	The values in each cell indicate MEAN (SD). The target messages with positive or negative typefaces achieved higher positive or negative ratings. Here, a higher score means more positive, a lower score means more negative. The values with an ** mark indicate a statistically significant difference at the .01 level.	63
5.3	The results of Task 2. The values in each cell indicate MEAN (SD). The participants reported higher valence ratings with positive or negative typeface than with the neutral typeface. The values with * mark indicate a statistically significant difference at the .05 level.	64
5.4	Vignettes for the role-playing test. The line (c) has variations in the typeface used. . .	65
6.1	Factor analysis result. The column α shows Cronbach's alpha.	78
6.2	The mean, standard deviation, maximum, and minimum of the ratings across all pairs by each group. We can see that font-image pairs designed by professional designers received the highest score. Next came System, Novice, and Random, respectively. We examined maximum and minimum scores of each task and found that Ours has the most stable performance.	82
6.3	Examples of the search results given image query (Restaurant, Figure 6.8a). We list font names and creators of each font as: 1: Mops by Uwe Borchert, 2: Hot Pizza by Dennis Ludlow (maddhatter_dl@yahoo.com), 3: green piloww by Billy Arget (billyargel@gmail.com), 4: Croissant One Regular by Tipo, 5: Xperience Pasta by Peax Webdesign, 6: Frijole by Font Diner, 7: Dreamwish by Starlight Fonts, Lauren C. Brown (lauren@ork.net), 8: Vanessas Valentine by bythebutterfly, Vanessa (BYTHEBUTTERFLY@GMAIL.COM), 9: Circle Of Love by cutieFont (cutiefont@gmail.com), 10: Snappy Service NF Regular by Nick Curtis. Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.	86
6.4	Examples of the search results given image query (Halloween, Figure 6.8b). We list font names and creators of each font as: 1: Unquiet Spirits by Sinister Fonts, 2: Hantu Kom Kom by Haslinda Adnan (kakalin2001@yahoo.com), 3: Haunted Eyes Regular by Misti's Fonts (mistifonts.com), 4: Jasper Solid (BRK) by AEnigma (kentpw@norwich.net), 5: JMH CRYPT Regular by Joorge Moron (joorgemoron@gmail.com), 6: Bloodytronic by Brain Eaters Font Co. (info@BrainEaters.com), 7: Raven Song Regular by Sinister Fonts, 8: Metal Macabre by BoltCutterDesign, 9: Casper by DJ-JohnnyRka, 10: Phantom Fingers Regular by Sinister Fonts. Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.	87

6.5	Examples of the search results given image query (Baby, Figure 6.8c). We list font names and creators of each font as: 1: Sweet Smile by cutieFont (cutiefont@gmail.com), 2: Sunshine Kiddy Font by Merethe Liljedahl, Lime (https://plus.google.com/114621834783881488812), 3: Fontdinerdotcom Huggable by Font Diner (http://www.fontdiner.com), 4: Fontdinerdotcom Luvable by Font Diner (http://www.fontdiner.com), 5: Baby Lexi Medium by bythebutterfly, Vanessa (BYTHEBUTTERFLY@GMAIL.COM), 6: Toyland NF Regular by Nick Curtis, 7: Unicorns are Awesome Regular by Misti's Fonts (mistifonts.com), 8: Addis Ababa by Kimberly Geswein (gesweinfamily@gmail.com), 9: KG Only*Hope by Kimberly Geswein (gesweinfamily@gmail.com), 10: The Happy Giraffe by Misti's Fonts (mistifonts.com). Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.	88
6.6	Examples of the search results given image query (Restaurant, Figure 6.8a). We list font names by rank. 1: PopJoyStd-B, 2: RodinWanpakuPro-DB, 3: RodinWanpakuPro-B, 4: RodinWanpakuPro-M, 5: TsukuARdGothicStd-E, 6: TsukuARdGothicStd-B, 7: TsukuARdGothicStd-R, 8: TsukuARdGothicStd-D, 9: TsukuARdGothicStd-L, 10: TsukuARdGothicStd-M.	89
6.7	Examples of the search results given image query (Halloween, Figure 6.8b). We list font names by rank. 1: ManyoKoinLargeStd-B, 2: ManyoKoinStd-B, 3: ComicMystery Std-DB, 4: MysteryStd-DB, 5: PopFuryStd-B, 6: TsukuBOldMinPr6-R6 SlumpStd-DB, 7: SlumpStd-DB, 8: arcStd-R, 9: RodinHappyPro-EB, 10: RodinHappyPro-UB.	90
6.8	Examples of the search results given image query (Baby, Figure 6.8c). We list font names by rank. 1: <i>UDKakugo_LLargePr6 – EL</i> , 2: <i>UDKakugo_ssmallPr6 – EL</i> , 3: <i>UDKakugo_LLargePr6 – UL</i> , 4: <i>UDKakugo_ssmallPr6 – UL</i> , 5: <i>TsukuMinPr5 – B</i> , 6: <i>TsukuMinPr6 – RB</i> , 7: <i>TsukuMinPr6 – D</i> , 8: <i>TsukuMinPr6 – LB</i> , 9: <i>TsukuMinPr6 – L</i> , 10: <i>TsukuMinPr6 – M</i>	91
6.9	Average scores of movie poster tasks. The score ranges from -2 (worst) to 2 (best).	94
6.10	Max scores of each movie poster task. The score ranges from -2 (worst) to 2 (best).	94
7.1	Examples of reconstruction given a single word and on following the pathway illustrated in Figure 7.4a.	107
7.2	The top five images obtained given an image input and on following the pathway illustrated in Figure 7.4b. All retrieved images look similar but are rendered by different font file.	108
7.3	Example of filling-in missing modality (text) given a single modality (image) on following the pathway illustrated in Figure 7.4d.	108
7.4	Example of multimodal output from multimodal input on following the pathway illustrated in Figure 7.4e.	111

7.5	Retrieval performance (<i>Recall@K</i>) comparison of original tag set ORI and randomly removed original tag set ORI-1, ORI-2 with generated tag set GEN, GEN-1, and GEN-2. #Tags indicates the average number of tags in a tag set.	114
7.6	Retrieval performance (<i>Recall@K</i>) comparison of adding random generated tags (RAN5/RAN10) with adding system generated tags (GEN5/GEN10).	115

Chapter 1

Introduction

Graphic elements, such as images and fonts, are good visual communication media. These media convey a variety of information, and they affect one's emotional state. A photo of sunset creates a feeling of warmth, and a text with the font Times New Roman would look professional. The goal of this thesis is to model systems that understand affective signals in image and font and to propose applications where the analyzed affective signals can be used.



Fig. 1.1 Two images that are the same in content (house, sky, grassland), but feel different.

1.1 Affective Signals in Image and Font

Affective Computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena [119]. The definition of the term *affect* is described as *subjectively experienced emotion* [109], and found that it subsumes emotions, feelings, and sentiments [49]. Figure 1.1 shows an example of what affect can be found in image and font. People would agree to annotate the two image Figure 1.1a and 1.1b with house, sky, and grassland. But they would feel different in both images even though the two images depict the same objects.

Figure 1.2 shows two different notes. Even though the same text message is written in the both notes, we may feel different depending on which font is used. Figure 1.2a sounds like affectionate but Figure 1.2b sounds like threatening. As seen from the above examples, contents such as category of the depicted objects and written text are one of the important components of image and visual text, but other cues such as color of the image and font of the visual text contribute to subjectively experienced emotion significantly.

Thanks to the recent advances in Deep Neural Networks, multimedia analysis have been developed. Especially large benchmark datasets accelerate this development such as ImageNet [87], MSCOCO [99], Speech [67], Movie [60], and Twitter [164]. These dataset provide images, speech data and a bunch of text with associated ground truth labels. With these benchmark datasets, many researches have focused on predicting the accurate label [87], descriptions [99] and so on. Leveraged by the improvement in recognition, high-performance retrieval systems have been proposed [151]. Meanwhile, some researchers have been concentrated on affect signals in multimedia. They say not only visually (or audibly) detectable object, but also emotion or feelings evokes in human receivers have a large impact on our interaction with systems, and further contribute to human-human interaction [139]. There are several benchmarks in affective computing. Multilingual Visual Sentiment Ontology (MVSVO) enabled us to understand images with adjective-noun pair (ANP), which is considered emotional descriptions than typical object categories [19]. Music-sentiment dataset [143] and text



Fig. 1.2 Two memos that are the same in text, but feel different.

sentiment mining dataset [162] also enabled researchers to analyze affect in multimedia and propose a new system. The development of affective system have satisfied users who think whether an object exist or not is not the only query for searching images. Sentiment understanding in movie review [115], and restaurant review [75] largely contributed to better recommendation results to users.

During the past decade, significant amount of User-generated content (UGC) have been generated and shared via on-line community such as Youtube, Pinterest and Dribbble. Users explores an enormous dataset to find the right image and the font that matches well with their intent of the contents. Especially, in graphic design process, the keywords used in searching are not limited to category of the object, but include affective keywords that explains emotions or feelings.

Motivated by these observations, we aim at modeling systems that understand affective signals in image and font and proposing applications where the analyzed affective signals can be used. In this thesis, we explore the answers to the following research questions: (1) How to model a system that predicts affects in images and fonts without a large dataset to learn (2) How to improve user engagement in searching with ambiguous and noisy dataset?

1.2 Advances in Affective Computing

1.2.1 Two perspectives on affective computing: Cognitive vs Interactional approach

From the 19th century, the interest of emotion began to start. However, the notion of emotion at the time — it is far from the science pursuing physically grounded and highly controlled space at the time, delayed the development of the research [44]. In the mid of 90s, against the common notion of emotion, Picard introduced emotion to scientific inquiry by making it scientific supported by cognitivist view [119]. The main idea of the paper is to give machine the abilities — recognizing and understanding one's emotional state, and expressing emotion. Based on this theory, most of the studies have been focused on modeling machines that mimic human. For example, they mainly detected visually-observable expressions (e.g., body gesture [125] and facial expression [101]) or acoustically-audible expressions (e.g., tone of the voice [132]). Based on the emotional categories such as Ekman's six basic emotions, they modeled systems that classify the emotional categories from given human facial expressions or tones of the voice. Multimodal methods that integrates two (or more) modalities such as facial expression and tones of the voice achieved improvement in emotion

recognition. These cognitive-based systems see affect as a measurable biological fact, and have been evaluated in terms of accuracy.

Alternative perspective on emotion is, the interactional approach. Rather than seeing the emotion as objectively classifiable, this view lets the system to see the emotion as ambiguous, subjective, and sensitive to context. It focuses on studying emotions as experience not accuracy. The most representative research in this view is to predict affects that a given object would create such as sentiment analysis on a writing such as blog post [24], review [75] and emotional state prediction of movie viewer [42]. This pragmatism based systems interpret emotions as they are made in interaction, and value practical usage of the understanding in our daily interactions in the real world. It evaluates system in terms of usefulness.

This thesis is based on the latter view, interactional approach. We see the computer as a medium that communicate affects between humans, not a machine that acts like human.

1.2.2 Scale of affects

There are two emotion classification models in general: categorical emotion states (CES) and dimensional emotion space (DES). CES was developed to define basic emotions based on the hypothesis that a basic emotion is biologically distinct from others. One of the most frequently exploited definitions was proposed by Ekman [45], who defined six basic emotions: happiness, sadness, fear, anger, surprise, and disgust. However, the word-based emotion category model have been criticized for lack of consideration on cultural differences [156]. For this reason, there was a movement to analyze emotions on continuous space rather than discrete categories. Dimensional Emotion Space (DES) is based on neurophysiology and defines emotions in two (valence-arousal) or three (pleasure-arousal-dominance) dimensions [133, 121]. Russell proposed circumflex model of emotion [129], which associates emotion terms with combinations of different intensities of valence and arousal. This work integrated both categorical and dimensional theory. Another integrated model is Plutchik's Wheel of Emotions [120]. It not only consists of 24 basic emotions, such as joy, anger, and anticipation but also explains emotion continuously in terms of intensity. This model can explain the degree of similarity among basic emotions. There also exists a categorical view on dimensional approach that determine the polarity of the valence (i.e., positive and negative) [36]. Though it limits the explainability, it has been used for many studies because of its simplicity.

Not only limited to emotion, there is another semantic space model for other affects, such as impression. Kobayashi devised color image scale [5], which associates color combinations with 180 affective words (e.g., warm, cool, hard, soft, light, heavy, fresh, masculine, feminine, classic, modern, active, and clean) and plots the words on the two dimensional semantic space consisting of warm-cool axis and soft-hard axis. More recently, with the advances in natural language processing, statistical models have been proposed to find latent topic in a set of documents [105, 117]. Word embedding such as Word2Vec [105] enabled calculating distance between words by numerical representation of words, so called word vector. Each element of the word vector is regarded as a latent variable. Word

embedding also allowed us to interpret the meaning of words using high dimensional continuous space. But unlike VAD space, it is difficult to explain each variable or axis.

1.3 Research Challenges

To understand affects in image and font, and make them meaningful to users, we encounter several challenges. Firstly, there is no ground-truth to predict feelings. We can clearly describe visually observable objects, but it is difficult to describe abstract feelings in a given image or font with a few words clearly. Second, it is challenging to apply interpreted affect information directly to actual system. The semantic gap problem would occur in actual interaction between system and human. Moreover, many systems have neglected affects because diverse and open-ended interpretation are obstacles for an efficient system with high accuracy [44]. Finally, font affect understanding has not been addressed in computer science. Many researchers in marketing field studied the effect of fonts in advertisement, but, few researchers studied emotional effects of font.

1.4 Objective and Contributions

This thesis focuses on the image and font. The main objective of this thesis is to understand affects in image and font, and propose affect-aware applications. The primary contributions of this thesis are summarized as follow:

Image impression retrieval: Conventional image retrieval systems ask users to input query by text.

However, it is not always easy for users to convert their intention into verbal representations. In Chapter 3, we propose an interactive retrieval system based on yes-no questions for image impression retrieval. We modeled a system that interprets images with impression words such as fresh and modern. Then, we introduce a yes-no question based querying method and a feedback interface to support users querying. From the user test, we showed that our system brings satisfactory results to users in case where the proper text querying is difficult.

Font emotion understanding: Different fonts create different experience. This ability of font have been utilized and studied in marketing and branding strategies (e.g., powerfulness of logo). However, there are few studies about font as an emotional modulator. In Chapter 4, we demonstrate the effect of fonts on viewer's emotional state by two experimental studies — explicit and implicit testing. In explicit testing, we measure the response to fonts using the assessment sheet which asks readers the feeling in the font directly. In implicit testing, we measure unconscious response to fonts using spontaneous speeches that elicited by fonts of written text. The series of studies show the potential use of font for emotional representation such as happiness and anger.

Font communication on mobile messenger: Instant messaging is a popular form of text-based communication. However, text-based messaging lacks the ability to communicate nonverbal information such as that conveyed through facial expressions and voice tones. In Chapter 5, we propose Emotype, a mobile messenger application prototype that enables users to change the font of a mobile messenger message to convey certain emotions. In user test, we demonstrate the feasibility of fonts for communicating emotions with a survey study, and then explore the unique feature of font that is different from other ways for expressing emotion by qualitative user study.

Font search by image: One of the important aspects in graphic design is choosing the font of the caption that matches aesthetically the associated image. In Chapter 6, we present two font search systems that enable users to use images as queries - (1) query by image impressions based on color study and (2) query by image contents based on concept analysis. Instead of matching font and image directly, we mapped both image and font to color-based semantic space or concept-based semantic space. Our evaluation results show that the recommended fonts scored better than other comparisons and provides competing results with the ones chosen by experienced graphic designers.

Creativity support in graphic design: Inspiration plays an important role in the creative process. By getting inspired, we can reach unexpected but useful ideas. Inspiration, generally, comes to us when we interact with external interventions. In Chapter 7, we present a framework that assist users' interactions in font search with unexpected but useful concepts generated by multimodal learning. By examining the results of the model that change with various inputs, we observed that the model produces promising results that appeared to be useful for inspiring users.

The remainder of this thesis is organized as follows. Image understanding study has been well established, but the effect of font has been studied only in a specific area such as marketing and never been examined interdisciplinary. In Chapter 2, we review studies on font in various fields over the centuries. In Chapter 3, we focus on image impression mapping on semantic space and then introduce two querying method, yes-no questions, and feedback slider. Chapter 4 presents the potential use of font for emotional representation by two experimental studies — explicit and implicit testing. Based on the study in Chapter 4, we propose an instant messaging applications that enables users to change the font of their message to convey certain emotion in Chapter 5. In Chapter 6, we propose two font search systems that enable users to use images as queries. In Chapter 7, we present a framework that support creative activity such as graphic design by providing unexpected but useful concepts to users. Finally, we conclude this thesis and suggest future direction in Chapter 8.

Chapter 2

Related Works: Research and Development on Font over the Centuries

There are now a large number of fonts in a diverse range of styles and each font creates a unique impression that affects viewers' emotions. To understand the properties of fonts, this chapter introduces how fonts are diversified along with development of technology over a long history, and then showcases a series of studies on font in computer science and related fields in recent decades.

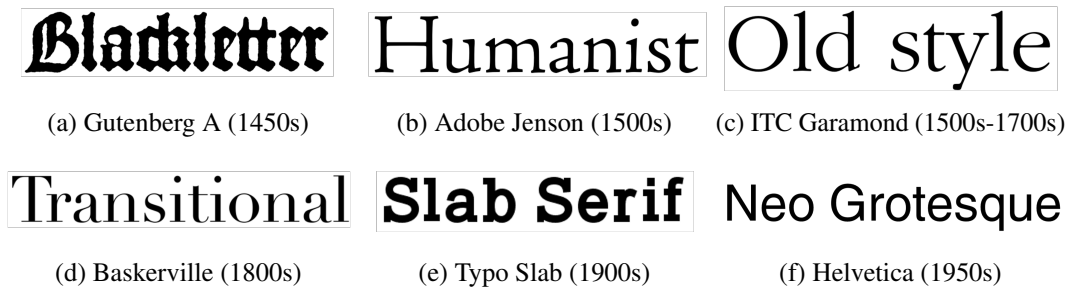


Fig. 2.1 Font development over the centuries

2.1 History of Font

Initial type of font Font is, currently regarded as an aesthetic medium, but it was initially devised with respect to printing technique. The history of the early font has a close relationship with printing. Since books were produced by hand before the invention of printing, it was not easy for the general public to access the books. However, with the invention of Gutenberg's printing process, the initial concept of font was created. In the printing age, the word *a font* refers to a set of metal type that has a particular size and style. The font were made to mimic handwriting at that time — blackletter type styles that have tall, narrow letters and sharp angular lines (Figure 2.1a).

Humanist In the 15th century, inspired by Italian humanist scholars' writing style, a new style of font *Humanist* developed. Especially, Poggio Bracciolini was famous for his beautiful handwriting, and it became the origin of the modern Roman style such as *Times New Romans* (Figure 2.1b). Humanist type became more popularized by Nicholas Jenson who is printer in Venice.

Old style Until the early Renaissance where metal type of font was usually made with reference to handwriting, font was completely calligraphic in nature. Between the 1500s and the 1700s, however, a style called *Old style* developed (Figure 2.1c). This new style is no longer a complete imitation of handwriting, but legibility has begun to be considered. Compared to the *Humanist*, the glyph has more straight cross bar, which looks more upright.

Transitional In the 1700s, *Transitional* were developed. It is more upright than *Old style*, so that the slope of character is perfectly vertical. The difference in the thickness of the stroke so called contrast began to be emphasized due to produce more elegant look font. Didone is a representative font of this age that shows high contrast (Figure 2.1d).

Slab Serif In the 18th century, European colonists discovered Egyptian Civilization. The enthusiasm for Egyptian culture, especially the architectures, inspired the production of *Slab serif*.

Early Sans Serif - Grotesque Influenced by *Transitional* and *Slab serif* styles, Sans serif began to develop officially. In fact, the letter without serif (Sans serif) has existed for a long time. However, because it was not considered to be sophisticated type, metal type of font was not made. As we can notice from the name of early Sans serif, *Grotesque*, it was not regarded as an elegant type to the public. Nevertheless, thanks to its characteristic that is clear from the distance, it was actively used in the title of posters.

Neo Grotesque In 1950s, *Neo Grotesque* is developed by Swiss designers who have the motto of simple but functional design so called *universal design* — it is the educational ideology of Bauhaus (Figure 2.1f). They tried to produce a type that has a neutral meaning, thus it is used in the communication where efficiency is important. However, since postmodernists began to criticize the triteness of the font, the experimental typographic, which influence the current fonts, appeared.

Contemporary movement Over the centuries, the styles of font have diversified. Especially, the number of new styles that have been created in recent decades is thousands of times larger than the number of styles created over the centuries. The main reason for the rapid increase in styles of font is that personal computers become a commonplace. As personal computers become more commonplace, printing paradigm has completely changed. The word *font* no longer only indicates metal typesetting but an electronic data file that renders any characters which share the same style. In other words, people did not have to produce metal typesetting that takes a huge amount of time, and this contributes to the growth of font market. Other major causes of the growth of market is the development of CPU performance. Because even complex shape of font can be rendered easily, diverse range of styles have created.

Now, the font market is getting bigger and bigger, over 170,000 fonts available on a single community. Since technical problems has been addressed such as required resources for rendering font, researchers began to study effect of font — readability, personality of fonts. The remainder of the section, we will introduce researches on effect of font, and showcases a series of studies on font in computer science and related field in recent decades.

2.2 Effect of Font

2.2.1 Readability of written text

As the visual appearances of font become diverse, the readability of each font in terms of communication via written text began to study so that information loss can be minimized [128][12][7]. The most actively studied factor affecting readability is serif, and it is still debated. Serif indicates the small details at the end of strokes. Fonts with serif have been shown to improve legibility, increase reading speed, and decrease fatigue. Rubinstein suggested that serif fonts may provide a cue to the location of

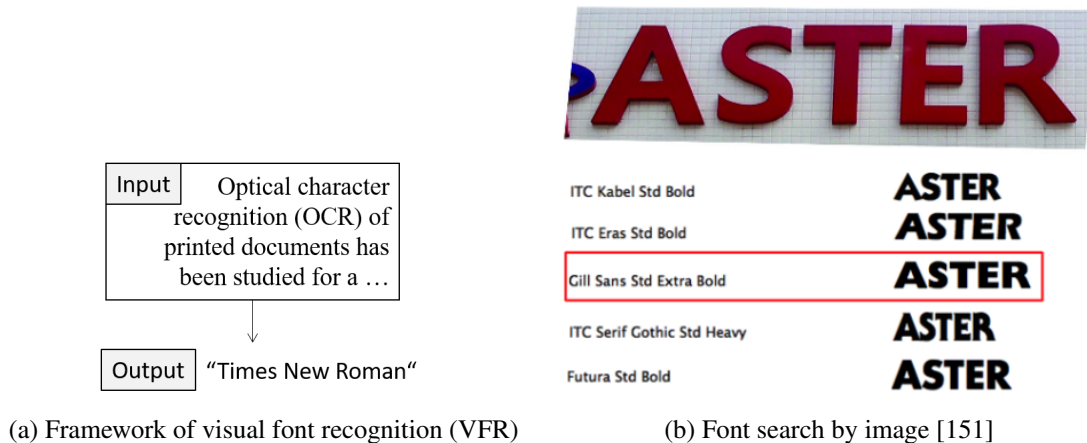


Fig. 2.2 Visual font recognition

stroke ends, which contribute to legibility [128]. However, some researchers have reported that no significant differences exist between serif and sans serif fonts in terms of legibility [18, 12].

Actually, there are many factors that affect the readability. There are studies that it is influenced by the display environment [148, 7]. The arrangement of text, e.g., kerning, line spacing, and letter spacing affect readability as well [127].

2.2.2 Personality of written text

As the font market grows, fonts are acknowledged as one of the most effective marketing tools for communicating the message. To utilize this effect of fonts, designers and marketing strategists studied how their audience may perceive fonts in use [9, 79]. The unique impression that each font may create is also known as *the personality of the font*. Shaikh et al. investigated the relationships between personality traits and preference for use of certain fonts [135]. They suggested that sans serif fonts are effective for website text or email, but serif fonts are more effective for business documents. Li and Suen examined the personalities of 24 typefaces and grouped them into four categories—directness, gentleness, cheerfulness, and fearfulness—based on survey data [96]. Amare and Manning explained why specific typeface features elicit certain emotional responses, thereby supporting the findings of previous studies based on empirical user evaluations [9].

Although the above described works focused on discovering and exploiting the personality of fonts, there is no consideration of exploring emotional influence of fonts such as happiness and anger.

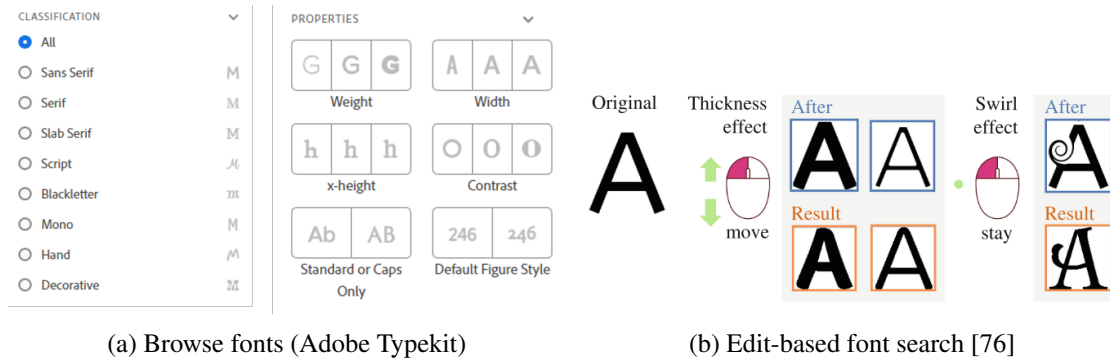


Fig. 2.3 Font search with clear user preference

2.3 Font in Computing

2.3.1 Visual font recognition

Optical character recognition (OCR) of printed documents has been studied for a long time since the beginning of the computer vision community. As shapes of fonts have become more diversified, including handwritten types, visual font recognition (VFR) has begun to make a mark [175] (Figure 2.2a). The initial motivations of VFR study were, (1) to recognize structures of document such as title and paragraph (2) to reprint the document, and (3) to improve the performance of OCR regardless of the font used [91]. Recently, the motivation of VFR has shifted with the growths of font market and user-generated content (UGC) market. To generate digital contents to share, people explore an enormous dataset to find the right font that matches well with their intent. Therefore, the recent motivation of VFR study is to allow users to search font what they want to find. This movement connotes that, people are well aware of the effects of fonts such as readability and personality.

Timely, technological advancement has enabled people to access large font datasets. The initial large-scale VFR study achieved 72.50% of top 1 accuracy with 2,420 classes with local image descriptors (e.g., SIFT or LBP) [23]. High-cost machine learning algorithms such as convolutional neural networks (CNNs) have been utilized to solve large-scale VFR problems. Through the use of deep CNNs, the classification performance obtained significant improvement so that it is applicable to a real-world image such as a signboard [154] (Figure 2.2b).

2.3.2 Font search

The most common way for searching font is to use terminologies that are used to describe the visual features of fonts, such as *serif* and *bold*. The typographic-specific dictionary has been developed [92], and it is still actively used on font web communities [38, 2, Adobe] (Figure 2.3a). Font categories such *Blackletter* and *Handwriting* also can be a good filter criteria.

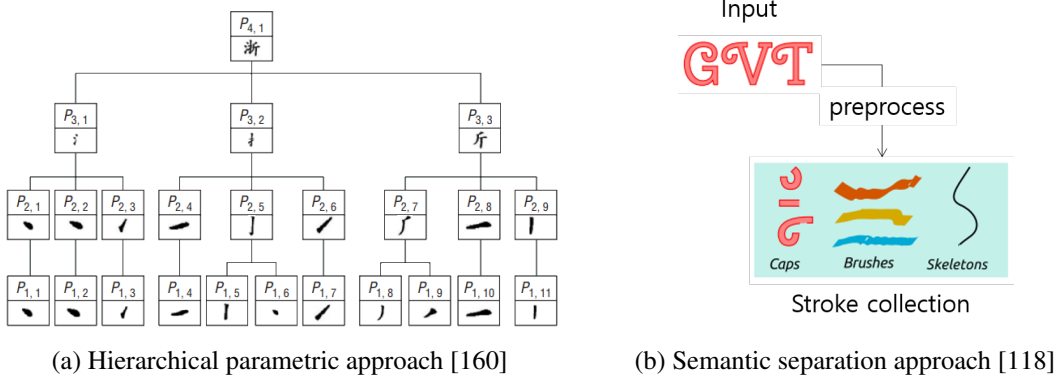
Thanks to the recent advances in computer vision, VFR has achieved remarkable performance. It enabled font search from real-world text images such as signboard for both English [154], and



(a) Exploration using high-level description [111]

(b) Exploration in 3D space [IDEO]

Fig. 2.4 Font search without clear user preference



(a) Hierarchical parametric approach [160]

(b) Semantic separation approach [118]

Fig. 2.5 Font shape learning by decomposing the shape of characters

Chinese fonts [70]. These systems are helpful when a user has a certain reference image for the typeface to be searched. VFR can be applied to sketch-based retrieval. It could be intuitive, but as like other sketch-based search methods, it is not good for the users who are not skilled. An interactive edit-based font search system is useful in this case. By adding simple deformation such as swirl effect to an original font, users can access to the goal font visually [76] (Figure 2.3b).

However, the above mentioned systems are only useful for people who have clear preference to font. To support these users, exploratory search scenario can be applied. Font Map [IDEO] is a new visualization tool that maps fonts in 3D space using visual similarity to allow users to explore in a new manner (Figure 2.4b). Campbell and Kautz [21] presented a manifold of fonts, and this enables users to explore new fonts are obtained by interpolating between existing fonts. Another exploratory search is using attributes. O’Donovan et al [111] proposed font search systems. With the system, users can explore fonts by using high-level description such as *dramatic* (Figure 2.4a), and organize fonts in a tree-based hierarchical menu.

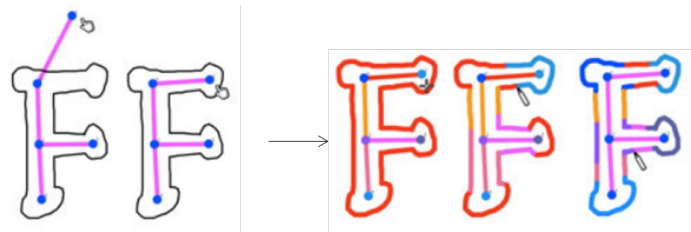


Fig. 2.6 Semi-automatic font generation that allow user intervention. Users need to adjust skeleton of the input glyph and select semantic parts [142]

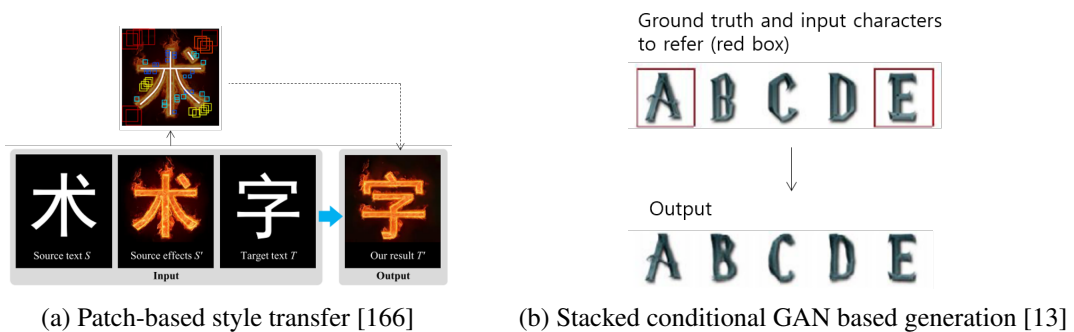
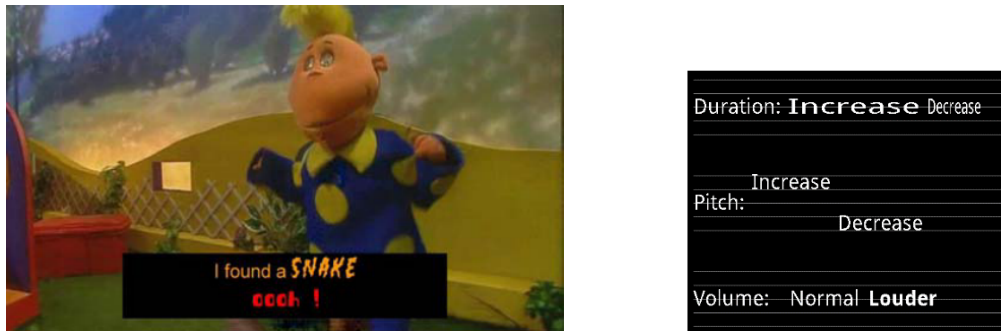


Fig. 2.7 Font styling

2.3.3 Font generation

Since fonts regarded as visual styles of written text, there have been attempts to separate content from style, and utilize each separated component, e.g., content for OCR, style for example-based automatic font generation. The very first font generation study proposed the idea that transfer a separated style from observed letters to new letters [144]. This study presented a multitask framework that learns both contents and styles using bilinear models. Generating Chinese font requires significant effort — it has more than 3,000 commonly used characters. For automatic Chinese calligraphy work generation [160], Xu et al proposed a hierarchical parametric approach that decomposes shapes of Chinese characters to represent a character combination of primitive strokes 2.5a. Similar to this work, Phan et al also decompose the shape of characters, but it differs in that it decompose the input glyphs into several semantic parts, such as caps, brushes, and skeletons [118] (Figure 2.5b). According to learned copying rules, the system assembles them and synthesizes missing glyphs. There are semi-automatic font generation methods that allow user intervention. Rapee et al presented a method that generate a new font from a user defined example [142] (Figure 2.6). In this system, a user draws an example outline of fonts. According to the user input, the system automatically computes the skeleton. The user can adjust the provided skeleton as necessary. By doing that, the system generates a set of characters that share the similar visual property of the example outline.

In addition to the glyphs generation, Font stylizing methods [13, 166] have proposed. Yang et al [166] explore the ornamented font generating problem. Given a source character, effect, and a target



(a) BBC's Tweenies television programme With Emo-tional Subtitles [112] (b) Typography Mapping to Prosodic Information [169]

Fig. 2.8 Applications that see font as a visual medium that conveys tones of voice

character, their patch-based style transfer method decorate the target character (Figure 2.7a). Recently, by leveraging recent generative adversarial networks (GAN), typeface glyph synthesizing from few samples by separating style from text image has developed [13, 172]. Azadi et al proposed a font style transfer using the conditional GAN architecture [13] (Figure 2.7b). Zi2zi [25] generates a new character by learning paired character images, but it requires a large set of paired data. In [172], a font style separating method, which consists of a style encoder, a content encoder, a mixer, and a decoder has proposed. The system separates style from a given set of characters, and then generates a character that have not been seen in the input.

2.3.4 Font as tones of voice

Font can be used not only as branding strategies but also as a medium for conveying emotions in text communication. Even if it communicates visually, there are many studies that see font as a visual medium that conveys tones of voice.

Document-to-audio (DtA) studies initially related typographic characteristics with characteristics of speech. These study said that typographical changes such as italic and bold induce a small change in the rhythm of speech [47] and further, contribute changes in pitch and speed of the voice [146]. This perspective was applied to speech synthesis for expressive speech synthesis to bold and italic fonts [147].

This characteristic of fonts also has been applied to interactive applications [112, 169]. An emotional subtitle system introduced fonts to provide audio cues for hearing-impaired community [112] (Figure 2.8a). This study tackled the problem that current subtitling systems only provide what is being said, and then proposed an emotional subtitle system that utilizes varying fonts for expressing emotional nuances. SpeechPlay [169] enables users to synthesize voice of written text by modifying the appearance of font using touch-based gestures (Figure 2.8b). In addition to the emotional effect of static text, the effect of kinetic typography — an moving text, has studied. Kinetic typography

emotionally influences the viewer through changes in animation, speed, and dynamics of written text and it has been applied to single-word message applications [82, 94].

Chapter 3

Image Impression Learning for Retrieval

We propose an interactive system based on yes-no questions for image impression retrieval. We propose two querying methods, a question generation method, yes-no question, and a feedback method. Conventional image search systems ask users to input queries by text. However, it is not always easy for users to convert their intention into verbal representations. Especially, the query generation becomes even more difficult when a user tries to find images with impression words due to its subjectivity. In addition, it is not guaranteed that the images are properly annotated with enough number and high quality of tags. To solve these problems, we propose a yes-no questions-based image retrieval system that can effectively narrow down the candidate images. We also provide a feedback interface in which users can do the fine tuning of weights of the impression words. We conducted experiments on image retrieval task with 117,866 images. The results showed that our system brings satisfactory results to users in case where the proper text querying is difficult.

3.1 Introduction

How can we succeed in image retrieval? From a system perspective, all the images should be annotated with high-quality of annotations. On the other hand, from a user perspective, users should design a query, which describe their intention well. Above all, bridging the semantic gap between the annotation and the user query is most important.

However, retrieving a proper image that matches well to user's intent is challenging for a system. The system not only needs to annotate images, but also needs to understand user query. Thanks to the recent advances in Deep Neural Networks, machines have become successful in image understanding such as image classification [87] and captioning [99]. But the classification or captioning only can be powered by large dataset with predefined classes or sentences to train. Therefore, a task with dataset consists of poorly annotated images and noisy tags is still difficult problem to solve. Consequently, an image retrieval system that considers impression of images and enables users to access with subjective query such as impressions or feelings of images as a keyword could not be developed.

In this chapter, motivated by these problems, we propose an interactive system based on yes-no questions for image impression retrieval. There are two contributions in this study. Firstly, we modeled a machine that interprets images with impression words. Most of the conventional image retrieval researches focused on understanding what kind of concepts are included in the image. However, image interpretation with concepts lacks subjective level of interpretation such as impressions on the image. This incomplete image understanding sometimes results in difficulty in finding satisfactory images. While an image pair with different contents can be easily distinguished by its annotation (Figures 3.1a and 3.1b), an image pair which have similar contents but different impressions is difficult to distinguish (Figures 3.1b and 3.1c). The modeling technique of this work is based on the color feature. Because the color feature is the most effective medium for expressing feelings and emotions [152], we conduct affect analysis based on the psychological color study which was designed for industrial fields, e.g., fashion, interior design and lifestyles. The second contribution is that we introduce a question-and-answer-based querying method into our system to support users querying. In general, to retrieve photos, users should submit detailed queries to the system. However, it is not an easy task for users to convert their intention into queries. In contrast to conventional querying system, our system gives a user questions which require yes-no answers. In this chapter, questions are automatically

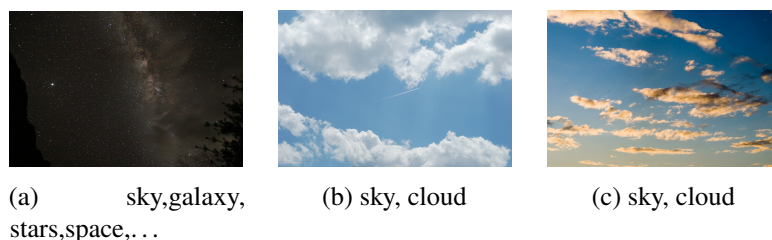


Fig. 3.1 An image pair (b) and (c) which have similar contents but different impressions is difficult to distinguish while (a) and (b) is easier to distinguish.

generated so that the candidate images will be efficiently narrowed down based on the distribution of the impression of the images. In addition to the querying support, we provide a feedback interface, which enables the user to update image results by changing impression factors. Finally, we investigate whether the affect analysis by the proposed system could be agreeable among users. Because affect analysis is highly subjective, perceptual differences may exist. This chapter explores answers to the following questions,

- How well our proposed method can retrieve images that the conventional text-based method is hard to access, and
- Whether semantic agreement exists between the system and users.

The remainder of this chapter is organized as follows. The impression analysis of images and user interaction studies on information retrieval are reviewed in Section 3.2. In Section 3.3, we introduce the user interface of our proposed system. We describe the methodology to map image on semantic space in Section 3.4 and querying method in Section 3.5. Then we show results of the experiments in Section 3.6. Finally, we analyze and discuss the retrieval result in the semantic space in Section 3.7.

3.2 Related Works

Many researches on affect and sentiment analysis have been conducted actively in the last decade. There are many theories for affect analysis of images, psychology and art theories are most frequently exploited [102]. Borth et al. devised ANP detectors to analyze sentiment from images by utilizing large-scale Adjective-Noun-Pair (ANP) on the web [19]. A state-of-the-art approach, such as multi-graph learning, is introduced for affect analysis [173]. These higher level of image analysis enable not only a system to understand the real-world, but also users to experience real-world on the system. Analyzing sentiment from actual user comments on social media can be utilized to understand real-world by system [130]. The affect understanding system could improve the quality of life, e.g., designing of magazine covers [78]. In many affect analysis works, the adjective words represent crucial elements for describing subjective representations.

However, Kato et al. observed that users' queries contain few adjectives in actual querying processing in information retrieval [81]. These researches promote to support users to express their intent with affective expressions. Many researches have investigated difficulties in querying. Pu reported that zero-hit searches often occur due to the difference between users' requests and system tag dataset provided in the image itself [122]. Šimko and Bieliková pointed out three drawbacks of the query-based search, expressiveness limits, keyword ambiguity and invisibility of information space [137]. To support users expressing their intent, various query modalities and processing methods have been developed, and the importance of Relevance Feedback (RF) have grown rapidly [39]. In this chapter, we propose an interactive system based on yes-no questions for image impression retrieval to solve the difficulties of querying.

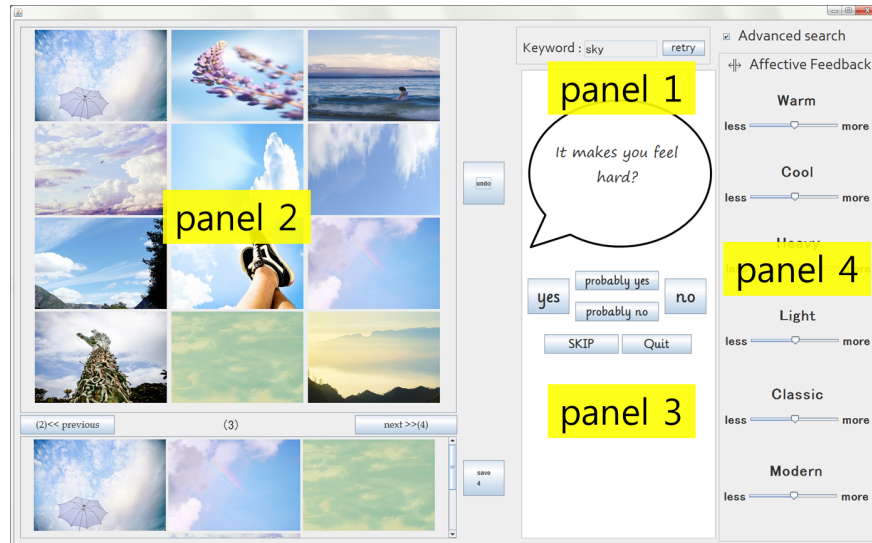


Fig. 3.2 Proposed image retrieval system

3.3 System Overview

Figure 3.2 shows the actual user interface of our system. A user inputs keywords, such as *Sky* (Figure 3.2-panel 1). Then the images tagged with *Sky* are randomly displayed in panel 2 (Figure 3.2). If a user want to refine the results, the system asks the user “Are you looking for warm feeling images?” and he/she can simply answer with *yes*, *probably yes*, *probably no*, *no* or *skip* (Figure 3.2-panel 3). We prepared a question template, “Are you looking for _____ feeling images?.” and filled the blank with predefined 13 impression words (*warm*, *cool*, *hard*, *soft*, *light*, *heavy*, *fresh*, *masculine*, *feminine*, *classic*, *modern*, *active*, and *clean*). Depending on user’s answer, the candidate images are reranked (Figure 3.2-panel 2). If necessary, he/she can give feedback to the system by adjusting feedback sliders (Figure 3.2-panel 4) to improve the results. For example, if the user is looking for images that feel warmer than candidate images, the user can adjust feedback slider by moving the *warm* slider closer to *more*.

3.4 Image Impression Analysis

When human perceive an image, the color, the direction of movement and the slope of a line are processed in human brain [11]. Among this information, the overall effect of the impressions is from the color features [153]. In this study, we exploit psychological color study [85] to construct subjective analysis model. We conduct image impression analysis. We used a term ‘impression words’ as a part of adjectives to describe impressions of an image have.

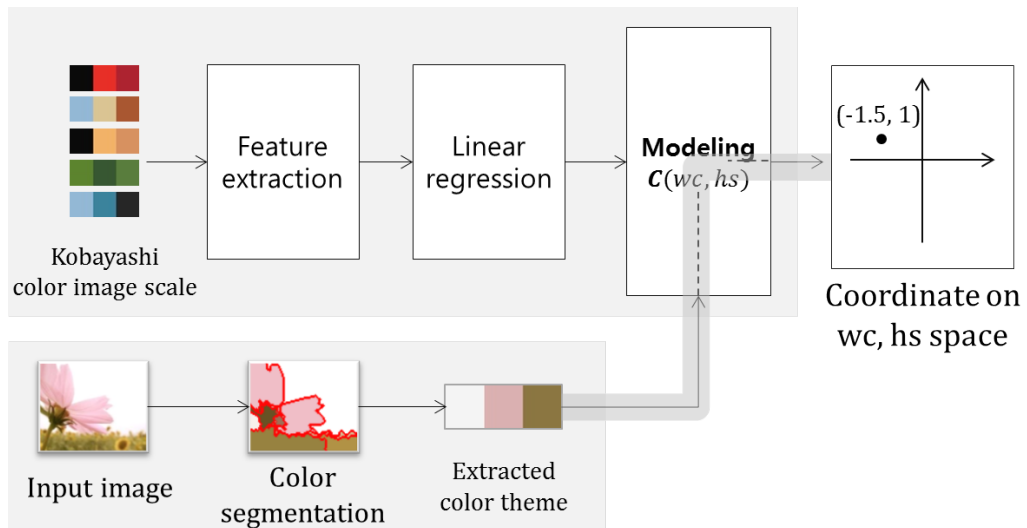


Fig. 3.3 Model training

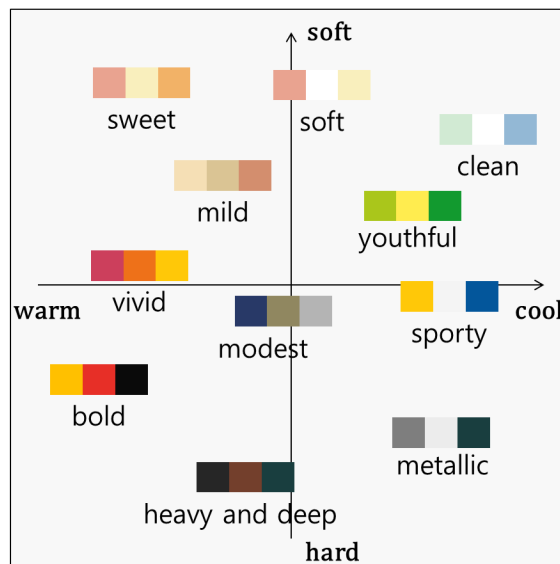


Fig. 3.4 Kobayashi color image scale on the Semantic Space

3.4.1 Kobayashi color image scale

Kobayashi designed color image scale [85], which associates color combinations with impression words for applications to industrial fields, e.g., fashion, interior design, lifestyles and so on. A single color combination consists of three colors of 130 colors of Practical Color Coordinate System (PCCS). He defined the relations between the color combinations and impression words. In his work, 1,170 sets of three-color combinations are organized from 130 basic colors. These combinations are matched to one of the 180 impression words which have a unique coordinate value on the two dimensional Semantic Space consisting of Warm-Cool axis and Soft-Hard axis, and the coordinate value is represented as (wc,sh) . Figure 3.4 shows examples of Color Image Scale in the Semantic Space. Because the distance between two color combinations means the difference of feelings, the combination also can be described as the other impression words by considering the distance. For example, a color combination associated with the word *soft* has a strong correlation with the word, *dreamy* than *serious*.

3.4.2 Impression word scale

We construct the impression word scale which is used to describe image in our system. In Kobayashi's work, it was shown that some groups of words share the similar impression. Therefore we reduce the number of word to use for our system. To understand the relation between the colors and words, some researchers have studied about what expressions are used to describe color [50, 110]. Ou, et al.[113] proposed 10 bipolar color emotion scales that are most frequently used in expressing colors with impression words (warm-cool, heavy-light, modern-classical, clean-dirty, active-passive, hard-soft, tense-relaxed, fresh-stale, masculine-feminine, like-dislike and intense-mild). We utilize the scale with little modifications suitable for our work. We exclude four words which have dominant cultural difference (tense, relaxed, like, dislike) and three words that are not defined in Kobayashi's work (dirty, passive, stale). Finally we set up 13 impression basic words for our system. The 13 words are *warm, cool, hard, soft, light, heavy, fresh, masculine, feminine, classic, modern, active, and clean*.

3.4.3 Model training

Despite ${}_{150}C_3=3,018,600$ color combinations could be generated by 150 basic colors, Kobayashi only provided the definitions only for 1,170 combinations. To predict the coordinates on the semantic space of a new color combinations that is not defined by kobayashi, we used 1,170 combinations and associated coordinates as ground truth to learn a model that predict coordinates of given new color combination (Figure 3.3).

We note a color combination as $C=(C_1,C_2,C_3)$. Here, C_i is a single color from 150 basic colors. The coordinates of C is represented by $C = (wc,hs)$. In addition to combination coordinates, Kobayashi also provided coordinates of a single color C_i . The coordinates of a single color C_i is

CIE-Lab space				
Color values	L	a	b	Calculation (e.g. L)
mean	L_{mean}	A_{mean}	B_{mean}	$L_{mean} = \sum_{i=1}^3 L_i/3$
maximum	L_{max}	A_{max}	B_{max}	$L_{max} = \max(L_1, L_2, L_3)$
minimum	L_{min}	A_{min}	B_{min}	$L_{min} = \min(L_1, L_2, L_3)$
max-min	$L_{max-min}$	$A_{max-min}$	$B_{max-min}$	$L_{max-min} = L_{max} - L_{min}$
variation	L_{var}	A_{var}	B_{var}	$L_{var} = \sum_{i=1}^3 (L_i - L_{mean})^2/3$
HSV space				
Color values	H	S	V	Calculation (e.g. H)
mean	H_{mean}	S_{mean}	V_{mean}	$H_{mean} = \sum_{i=1}^3 H_i/3$
maximum	H_{max}	S_{max}	V_{max}	$L_{max} = \max(H_1, H_2, H_3)$
minimum	H_{min}	S_{min}	V_{min}	$H_{min} = \min(H_1, H_2, H_3)$
max-min	$H_{max-min}$	$S_{max-min}$	$V_{max-min}$	$H_{max-min} = H_{max} - H_{min}$
variation	H_{var}	S_{var}	V_{var}	$H_{var} = \sum_{i=1}^3 (H_i - H_{mean})^2/3$

Table 3.1 30 Color Features (color values that defined by two space, CIELab and HSV)

represented by $C_i = (wc_i, hs_i)$. Using these color combinations, we defined 35-dimensions color feature vector CF on each 1,170 color combination $C=(C_1, C_2, C_3)$.

Firstly, we get the values L_i, A_i and B_i by CIELAB color space, and H_i, S_i and V_i by HSV color space for the color C_i . Table 3.1 shows defined 35 color features and corresponding formulation. In addition to color value based features, we additionally defined 5 more color features using coordinate values (Table 3.2). Here, the feature ED is defined as the emotion distance among the three colors combination C and it is calculated by summing of Euclidean distance between C_1^s and C_2^s , C_2^s and C_3^s and C_3^s and C_0^s . We denote the defined 35 color features as CF_k ($k=0, \dots, 34$).

Because we have 1,170 color combinations, we calculated 1,170 feature vectors and the set of vectors is learned by linear regression. We defined two linear regression models for each axis as following,

$$Warm - Cool : C(wc) = \sum_{k=0}^{34} F_k^{wc} \times CF_k,$$

$$Hard - Soft : C(hs) = \sum_{k=0}^{34} F_k^{hs} \times CF_k.$$

Here, F_k^{wc} and F_k^{hs} ($k=0, \dots, 34$) are parameters to learn that corresponding to 35-features CF_k . With learned parameters, we obtain the correlation coefficients between the training and the testing by 10-fold cross validation. They are 0.91 and 0.845 for Warm-Cool and Soft-Hard respectively.

Because the input of the model is a 3-color combination, we need to extract the three representative colors from given image. All the images in the dataset were segmented by using super-pixel generation method in [97] ($k = 10$, where k indicates the number of segmentation regions). Then we define that

Semantic space coordinates			
Coordinates	Warm-Cool (WC)	Hard-Soft (HS)	Calculation (e.g. WC)
mean	WC_{mean}	HS_{mean}	$WC_{mean} = \sum_{i=1}^3 wc_i / 3$
var	WC_{var}	HS_{var}	$WC_{var} = \sum_{i=1}^3 (wc_i - WC_{mean})^2 / 3$

Table 3.2 5 Color Features that calculated using semantic space coordinations

the colors of the three biggest areas are the 3-colors $C = (C_1, C_2, C_3)$. Then the coordinate value of an image is calculated using the above equations with learned parameters.

3.4.4 Impression score

By now, we showed how to estimate the position of an image on the semantic space. Although the coordinates implies the feeling of the image, it is not comprehensible to users. To describe images with intuitively, we assigned 13 impression scores to each image considering the estimated position. Because the distance between the two coordinates of two images on the semantic space means the degree of difference between the feelings of each image conveying, a score $P(imp | I)$ (or p_{imp}) of an image I given impression word imp is calculated as:

$$P(I_{imp}) = p_{imp} - \frac{1}{3\sqrt{2}}d_{imp} + 1$$

$$d_{imp} = \sqrt{(wc_{imp} - wc_I)^2 + (hs_{imp} - hs_I)^2}$$

Here, wc_I and hs_I indicate the coordinates of warm-cool and hard-soft axis of an image I respectively. wc_{imp} hs_{imp} is the coordinate value of one of the 13-impression words. Finally we can get the 13-dimensional score vector of an image. For example, $p_{warm} = 0.9$ means that we expect the given image I gives a highly warm feelings to users ($0 \leq p_{warm} \leq 1$).

3.5 Querying Support

Existing image search system asks users to input queries. However it is not easy task for users to design a proper query. To support query making, we propose two novel interactive querying methods, yes-no questions and slider feedback.

(1) Yes-no questions: With conventional image retrieval services, users should submit detailed and effective textual queries. Although they can describe what object concepts are included, it is not easy for users to describe their impressions or feeling into refined verbal representation. To support querying difficulties, we propose a novel querying method base on interactive yes-no questions. With the yes-no querying method, motivated by 20 questions game, users do not need to struggle to modify

their query. If a user asks for help to the system, the system questions, e.g., “Are you looking for warm feeling images?” and he/she simply answers yes-no.

(2) Slider feedback: Although subjective expressions could be agreeable among users, there still exist perceptual differences. To avoid these individual-variation-driven dis-satisfactory results, we support users to give feedback to the system, slider feedback, which enables users to update candidate images by controlling the slider intuitively.

3.5.1 Yes-no questions

Yes-no questions method is divided into three parts, system question, user answer and ranking. The system asks users questions which can be answered with yes-no. Therefore it is important for the system to choose the most effective questions among 13 impression words. After questioning, a user answers to the question. Not only *yes* or *no*, we also support fuzzy answers with low confidence, such as *probably yes*. After all, the images should be sorted by the score according to the question-answer pair.

Question generation

To choose the most effective questions from predefined 13 questions, the system refers to the standard deviations of each impression word’s probability score. For example, if all images have high *warm* score (or low low), the question “Are you looking for warm feeling images?” would make most images included (or excluded) as candidate images. Therefore the question made of *warm* (has low standard deviation) is not good for filtering images.

Images on the dataset are known by the system, the distribution of all images on each impression word can be calculated in advance. $STD_{set} = \{STD_{imp1}, STD_{imp2}, \dots, STD_{imp13}\}$. Then, an impression word that has more non-uniform distribution (the highest standard deviation) is chosen as m^{th} question Q_m . In actual retrieval system, we apply the querying method as fine-tuning method given user-designed query results. In other words, the system refers to the score of top k images matched to user-designed query to choose a question.

Processing five different type answers

Users can answer with five answers (a_m), *yes*, *no*, *probably*, *probably not* and *skip* given question Q_m . According to the user answer, the estimated score that each image to be chosen is updated. Here, $P(Q_m|yes)$ ($Q_m = imp_i$) stands for that the estimated probability of an image I will be chosen given a user answer *yes* to the question Q_m , and it is identical to $P(imp_i)$ which we have derived in Section 3.4.4. We have only the probability of *yes* given Q_m ($P(Q_m|yes)$) by our impression analysis model not for other possible user answers.

To support the five types of answer, all the possible question-answer pair scores are prepared before user image search. We introduce the information entropy to process the fuzzy answer from

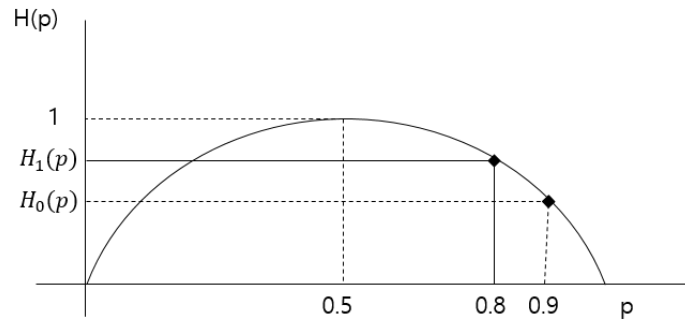


Fig. 3.5 Processing of fuzzy answer from user



Original score	p_0 (predicted value how much the image is "warm" by proposed model)	0.9	0.7
↓ by Q&A querying			
(ans)	Score of an image $P(warm? = ans image)$		
Yes	$p_{warm=yes} = p_0$	0.9	0.7
No	$p_{warm=no} = 1 - p_0$	0.1	0.3
Skip	$p_{warm=skip} = 0.5$	0.5	0.5
Probably	$p_{warm=p.yes} = 1/2(1 \pm (\sqrt{1 - H(p)}))$	0.7	0.6
Probably no	$p_{warm=p.no} = 1/2(1 \pm (\sqrt{1 - H(p)}))$	0.3	0.4

Fig. 3.6 Processing of fuzzy answer from user

users. Figure 3.6 illustrates this process. For example, if the user answers “probably yes” to the question “Are you looking for warm feeling images?”, the question-answer pair with low confidence should have a less effect on re-ranking than “yes/no”. Supposing that the score of an image I being warm is p_0 . An the score p_a is the updated score depending on the user answer, such as *probably yes*. We define that the uncertainty of user answer makes the entropy gets lower than before and higher than that of *yes*. Suppose the entropy, when the image gives *warm*, is $H(p_0)$. We assume the entropy is $H(p_a)$ when the image is probably warm, and $H(p_a)$ is higher than $H(p_0)$ (Figure 3.5). Here, we assume this process as $H(p_a) = 1/2(1 + H(p_0))$. p_a is obtained by solving this equation (See Figure 3.6 for each solution). In case of high confidence, such as *yes* and *no*, the updated probability p_a is equal to p_0 . The equation for answer, *probably yes* and *probably no*, makes the high score lower, and the low score higher both approaching to median value (0.5), where have highest uncertainty. Offline stage, we calculate all the scores of images for every question-answer pairs.

Reranking

If a question-answer is iterated, the number of question-answer pair increases. To rerank the set of images, size of N , given multi-question-answer queries, the score of an Image I_i given multi-question-answer pair Q_m and a_m is calculated by

$$P^*(I_i | Q_1, \dots, Q_n) = \frac{P(Q_{1|a_1} | I_i)P(Q_{2|a_2} | I_i) \dots P(Q_{n|a_n} | I_i)}{\sum_N P(Q_{1|a_1} | I_N) + P(Q_{2|a_2} | I_N) + \dots + P(Q_{n|a_n} | I_N)}.$$

Then all the image are sorted by the score.

3.5.2 Feedback slide

In addition to the question-and-answer-based system, we support fine tuning by feedback slide. Although impression words can describe the image well and there is consensus among people on a perceptual expression [122], the perceptual level differences among users still exist. In this stage, a user gives feedback to get more satisfactory images which correspond to the user’s intention. Impression words can be accompanied with comparative terms such as *less* or *more*, and we applied this property to our system. With our proposed method, a user controls sliders. The user can get more *fresh* images by moving *fresh* slider to *more* (Figure 3.7a). Because 13 impression words have correlations among them, if the user controls a single slider, the others are also adjusted in accordance with the variation of the slider. If a user moves the slider of an impression word s , for Δm , initial coordinate (wc_0, hs_0) will be changed tracing the line which connects (wc_s, hs_s) and (wc_0, hs_0) . Then we calculate the value of variation of each wc and hs coordinate by

$$(\Delta wc, \Delta hs) = (wc_0 \pm \frac{\Delta m}{\sqrt{g^2 + 1}}, hs_0 \pm \frac{\Delta m \cdot g}{\sqrt{g^2 + 1}}).$$

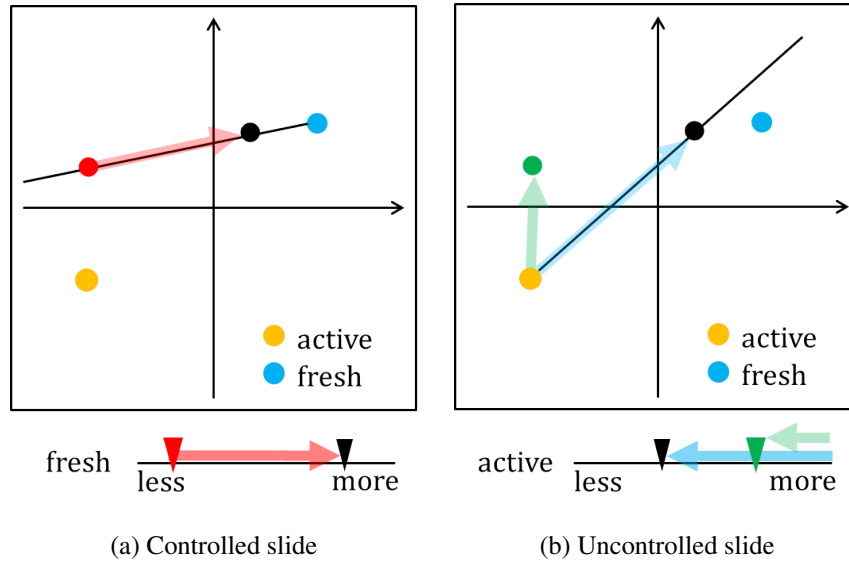


Fig. 3.7 Feedback slide

Here, $g = (wc_s - wc_0)/(hs_s - hs_0)$. Because the change of the coordinate affects the distance from the other 12 impression word coordinates, the slider of the other impression words are also changed, though the user controls a single word slider. Finally, the other sliders are adjusted in accordance with the variation of the slider (Figure 3.7b).

3.6 Experiment

In this section, we evaluate user satisfaction of the proposed system. We conduct user studies on image retrieval tasks with proposed method and conventional text-based system. We give users image search tasks. After all the tasks are finished, they review collected images and conduct the final selection. By comparing the number of images finally selected, we can evaluate the user satisfaction.

3.6.1 Task design and dataset

We collected images from Flickr¹, one of the most popular image hosting websites. It provides a list of popular tags, so we make use of the list to construct the image database. However many of tags are geographical place names. Therefore, we categorized popular tags into six classes and selected two keywords from each class for not biased task organization. The classes and keywords are Event (*Birthday, Wedding*), People (*Family, Friends*), Place (*California, NYC*), Activity (*Travel, Art*), Nature (*Beach, Sky*) and Season (*Snow, Summer*). Then we downloaded 10,000 images per keyword. Finally, we collected 117,866 images for the system evaluation (2,134 images are duplicated).

¹<https://www.flickr.com/>

Table 3.3 Experimental results showing the average number of positive-group images for each query. In most cases (nine-twelfth), the number of positive-group images collected by the proposed system is more than text-based system. It indicates that the proposed system was more satisfactory for the users.

-	Task	Proposed system	Text
T1	Quite Birthday	6.00	5.14
T2	Sweet and Dreamy Wedding	7.35	5.07
T3	Nostalgic Friends	5.78	6.92
T4	Dignified Family	5.21	5.42
T5	Lively California	6.79	6.64
T6	Free NYC	6.64	5.42
T7	Happy Travel	5.14	6.64
T8	Serious Art	6.93	4.5
T9	Wild Beach	6.42	6.14
T10	Mild Sky	7.00	6.29
T11	Pure and Simple Snow	7.14	6.29
T12	Subtle and Mysterious Summer	5.29	4.92

3.6.2 Task design

We generated 12 phrases by ourselves, which are composed of one noun (such as *Birthday* or *Travel*) and one adjective word (such as *Quite* and *Happy*) (Table 3.3). The 12 nouns are the same as the 12 keywords introduced in 3.6.1, and 12 affective words were selected from Kobayashi color image scale 180 words [85]. The 12 words are evenly distributed on the Semantic Space. Here, note that the 12 adjective words are not included in predefined 13 questions, thus these are independent from adjective words that are used in the 13 questions. Finally, we arranged highly subjective 12 phrases, e.g. *Sweet and Dreamy Wedding*. 14 persons participated in our experiment (2 female, 12 male, age of 22-28). We provided the participants for a task phrase, e.g. *Quite Birthday*. And the participant imagined how it looks like for a minute. Then, he/she conducted image retrieval tasks with two systems text-based system and proposed system.

We employed LUCENE², open source text-based search engine, as a conventional text querying system for comparison. In this system, for example, a user designed queries such as *lonely* by himself/herself to accomplish the *Quite Birthday* task. In case of the proposed system, he/she was not asked to input additional query but simply answered yes-no and tuned feedback sliders as necessary. Within 3 minutes, the user collected 10 images per task. Finally 20 images were collected by our system and the text search system. After all image retrieval tasks are finished, each user will have 20 images (or less for the duplicated images) in each task phrase. Then the user reviewed these 20 images (or less if any images are common for the both search) and classified images as positive-group, considered to correspond to their intention (Figure 3.8). The remainder is classified as negative-group.

²<https://lucene.apache.org/>



Fig. 3.8 Example images for the *Sweet and Dreamy Wedding* task (solid line images were collected by suggest system and broken line images were text-based system). Majority of positive images were collected by the proposed system.

3.6.3 Results

Table 3.3 shows the average number of positively marked images on each task with two different systems. In most cases, the average number of positive-group images collected by the proposed system is more than text-based system. It indicates that the proposed system is more satisfactory for the users.

However, the results of *Friends*, *Family* and *Travel* were not good. In the experiment, some users tended to judge the images with facial emotions in case of People class. Those users have a need to retrieve images by emotional words. However our system does not support the needs. So, the users were satisfied with text-based system, freer querying is available in People class. In case of *Happy Travel* task, all users were using *Happy* as a query which is one of the frequently used tags on the image dataset (112th of a total 104,846 tags). Eventually, it leads to high rate of matching results with conventional querying system. In conclusion, our system brings about satisfactory results to users in case where the proper text querying is difficult.

3.7 Discussion

In this section, we verify that the affect analysis model agrees with the color psychological model covered in this chapter. As we mentioned in the Section 3.4.3, we can get the coordinates of the processed image I , which can be plotted in the Semantic Space by the model (Figure 3.9). Because the Semantic Space includes the concept of impression meaning, the images plotted in the Semantic Space can be interpreted in impression meanings. Therefore, the distribution of images on the Semantic Space gives a general view of how different the prominent feelings of collected images by the proposed

system and text-based system. In this section, we analyze the distribution of images on the Semantic Space to answer following questions.

- How well the proposed method can retrieve images that conventional text-based method cannot access.
- Whether semantic agreement exists between the system and users.

3.7.1 Proposed vs Text-based

Figure 3.9 shows the image distribution of two tasks, *Sweet and Dreamy Wedding* and *Happy Travel*, on the Semantic Space. Figure 3.9a(1) shows the distribution of all images tagged with *wedding*. Figure 3.9a(2) shows the distribution of images selected by the text-based system, and Figure 3.9a(3) shows the distribution of images selected by the proposed system. The distribution of images tagged with *wedding* is spread in specific region (Figure 3.9a(1)). According to Kobayashi space [85], we can say that the images located in this region tend to be more *natural, elegant, chic, dandy, classic, formal*, and not to be *pretty, dynamic, romantic, clear, cool* and *casual*. The distribution of the images retrieved by the text-based approach seems almost similar to the total distribution (Figure 3.9a(2)). This observation implies that the feelings of images retrieved by the text-based system are not well differentiated from that of total images. This tendency can also be observed in the *Happy Travel* task as well.

However, the distribution of images collected by the proposed system has significant difference from both distributions of entire images and result of the text-based search. As shown in Figure 3.9a(3), we can observe that the distribution tends to be concentrated in middle of warm-cool axis and top of soft-hard axis. This region means the images are neutral in warm or cool, but highly soft. According to Kobayashi space [85], the region where images are concentrated is defined as *romantic, dreamy* and *peaceful* region. The meaning becomes closer to *Sweet and Dreamy* than that of text-based search. In case of *Happy Travel* task, the distribution of images collected by proposed system also has different distribution in comparison with entire images and text based search results. We can see that the distribution of images retrieved by the proposed system has different characteristics from that of the total image. From this result, we demonstrate that the proposed system provides images that are difficult to be accessed by the conventional text-based system.

3.7.2 Semantic agreement

We investigate whether the semantic agreement exists between the proposed system and users. Figure 3.10 shows the distribution of images selected by the proposed system only. Here, we skip showing the distribution of the text-based system and total images due to that these are very similar to Figure 3.9. In case of Figure 3.10 (T2: *Sweet and Dreamy wedding*), the square star mark (yellow) indicates the image of *Sweet and Dreamy* on the Semantic Space defined by Kobayashi (also our model). We can notice that the images collected by the proposed system are concentrated in this region. This tendency

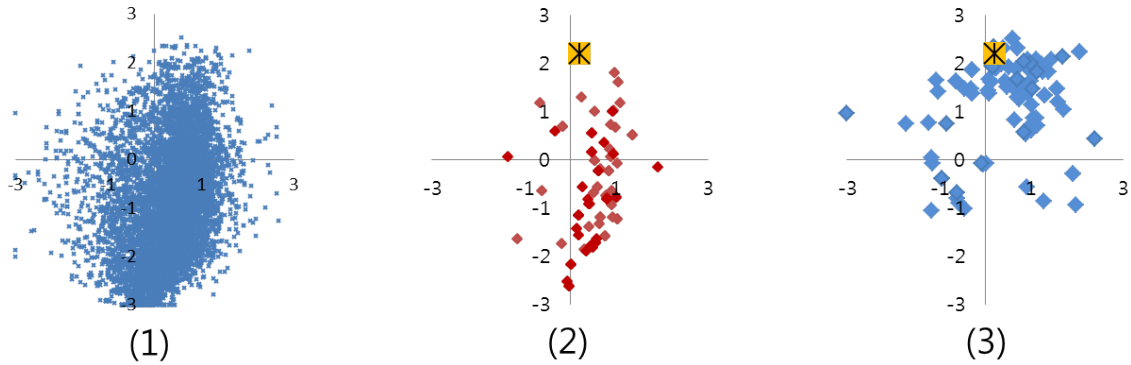
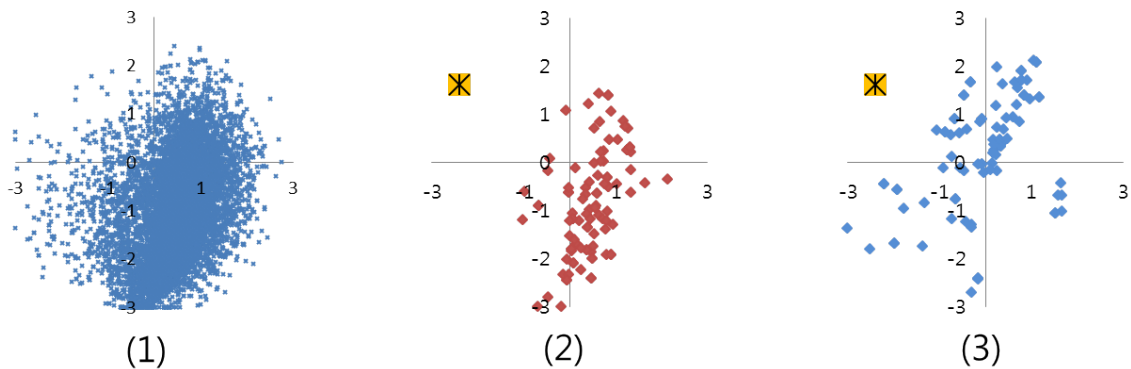
(a) The results of *Sweet and Dreamy Wedding*(b) The results of *Happy Travel*

Fig. 3.9 Image Distribution on the Semantic Space (x-axis: Warm-Cool, y-axis: Soft-Hard). (1) shows the distribution of all images tagged with a task noun, e.g., *wedding*. (2) shows the distribution of images selected by the text-based system. (3) shows the distribution of images selected by the proposed system. The distribution of (2) is similar to (1) but (3) is significantly different from (1) or (2). It implies that the images retrieved by the text-based system are not well differentiated from that of total images.

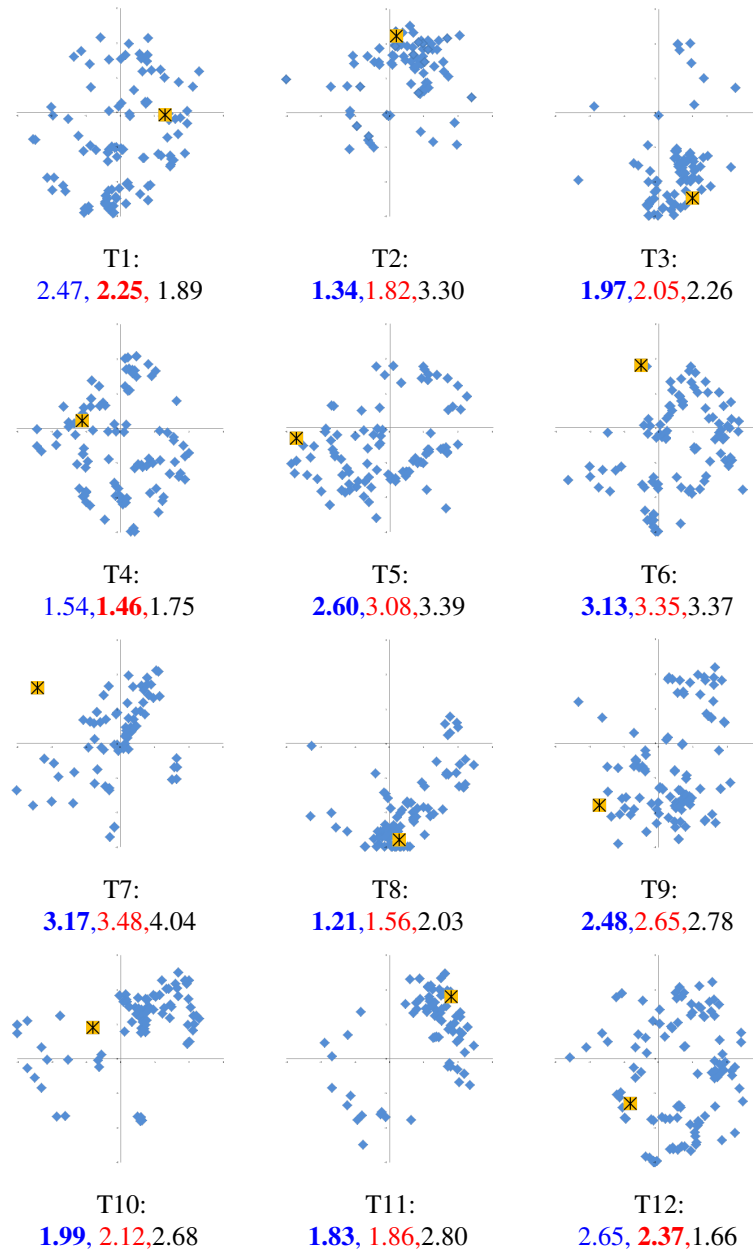


Fig. 3.10 Image Distribution on the Semantic Space (Proposed system), x-axis: warm-cool, y-axis: soft-hard. The average Euclidean distance is shortest with our system (blue colored) in most cases, and it implies that the images collected by proposed system has best consensus with Kobayashi model.

implies that not only semantic agreement exists among users, but also many of them agree with Kobayashi model. However, in the case of *Happy Travel* task, semantic image agreement between the users themselves, and between each user and the system is not very clear. For more detailed analysis, we calculated the average Euclidean distance from a task's adjective coordinates (square star mark) to the coordinates of images. For example, if the position of image *I* gets closer to the position of *Sweet and Dreamy*, we can say that the Euclidean distance gets smaller and the meaning of image *I* gets closer to *Sweet and Dreamy*. The calculation results are captioned with each subfigure. The calculation results with positive-group images collected by the proposed system is in blue colored, the text-based system is in red colored, and with entire images are in black colored. From this result, in most cases (except T1, T4 and T12), the average Euclidean distance is shortest with our system, and it implies that the images collected by proposed system has better consensus with Kobayashi model than both entire images and the text-based result images.

In conclusion, images collected by the proposed system are closest to the coordinates of task adjective on the Semantic Space in most cases. We observed that the distribution of images retrieved by the proposed system has significant difference compared to that of the total images. These results imply that impression analysis of images helped users to reach more relevant images. The images collected by the proposed system have a stronger tendency of concentrated distribution than the text-based search results. It implies that the affect analysis by proposed system could be agreeable among users. These results suggest that the analysis of affect can be taken advantage of for image retrieval.

3.8 Conclusions

We proposed an interactive system based on yes-no questions for impression image retrieval. Our system assists users to narrow down candidate images, which have similar contents but different impressions by yes-no questions and feedback slider. We conducted experiments on image retrieval task with large image dataset. By comparing with the text-based system, our system brings about higher satisfactory results to users in case where the proper text querying is difficult. Also, in comparison with distribution of images retrieved by text-based system, the impression analysis of images helped users to reach more relevant images. We also observed that users have similar visual imagery to highly subjective task and the proposed system based on color psychological model agreed among users well.

Chapter 4

Font Emotion Understanding: Measuring Explicit and Implicit Emotional Responses to Font

Font influences the received impression among viewers, and the ability of font have been utilized and studied in marketing and branding strategies (e.g., powerfulness of logo). However, there are few studies about font as an emotional modulator. In this chapter, we investigate the potential use of font for emotional representation such as happiness and anger. We constructed a dataset of various fonts and collected emotional responses to the fonts. To collect emotional responses, we conducted two experiments — explicit and implicit testings. In explicit testing, we requested participants to report emotions each font creates based on commonly used psychological emotion scale via crowd-sourcing. In implicit testing, we invited participants in site, and requested them to read out sentences written in different fonts. From the explicit testing, we determined that several visual characteristics of fonts have weak/strong emotional influences on viewers. From the implicit testing, we observed that speeches obtained by the fonts which have high-emotional influence are characterized by broader excursion range, longer speech duration, and higher pitch indicating the more emotional speech.

4.1 Introduction

Different fonts create different experience. Many researchers in marketing field was well aware of the influence of font [9]. They recognized the importance of font as communication media, and studied font effects such as impressions of advertised brands [63], memorability of advertisements [28]. However, compare to the studies that recognize fonts as branding strategies, there are few studies about font as an emotional modulator. In this chapter, we investigate the potential use of font for emotional representation such as happiness and anger. We have some indication that a font can provide emotional signals. For example (see Figure 4.1), compare to one of the commonly used font Arial, font with rounded cap evokes a positive feelings, and the sharp font evokes a negative feelings [33].

To demonstrate the effect of font on viewer's emotional state, and understand its characteristic as a emotional modulator, we designed two experimental studies — explicit and implicit testing. In explicit testing, we measure the response to fonts using the assessment sheet which asks readers the feelings in the font directly. By doing that, we expect to know constructed opinions about font among users such as whether emotional consensus toward a font exist or not. In other hands, implicit testing measures unconscious response to fonts. We investigate the font effect on reader's emotion during reading. In other words, despite the readers were not asked to express emotions, the written text with unusual font simultaneously affects the reader's emotion, and makes the reader to unconsciously produce more expressive speech than the commonly used font text. And then, we see whether implicit responses toward fonts are different from explicit response.

This chapter is divided into following sections:

- Font dataset : Construct a dataset including various fonts.
- Explicit testing : measuring the emotional responses to fonts using assessment sheet that request participants to report emotions each font creates directly.
- Implicit testing : measuring the emotional responses to fonts using spontaneous speeches that elicited by fonts of written text.

Firstly, we constructed a dataset with various fonts that have no problems for research purposes. Even though the shapes of fonts have diversified, many existing font dataset use a limited number

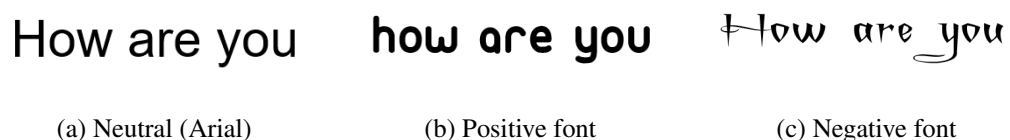


Fig. 4.1 Examples of a text varying fonts. Texts can appear to convey negative feelings or positive feelings depending on the font used.

of fonts. Recently, some studies constructed large-scale font dataset, but the majority are composed of only generally used fonts such as Times New Roman and Arial [111, 155]. To collect diversified responses to fonts, we develop a dataset of various fonts. Second, we collected quantitative user labels for each font based on commonly used psychological emotional scale via crowd-sourcing. By analyzing the user label data, we expect to see whether there exists an emotional consensus toward a font. Further, the collected emotional labels can be useful to see and demonstrate which font has a strong or weak emotional influence on the audience. Finally, we examine the effect of fonts on oral reading by providing prosodic analysis for speeches obtained from a different font stimulus. In addition to the analysis, we introduce the subjective evaluation on the obtained speeches to determine how well people recognize the emotion naturally induced by the fonts.

4.2 Related Works

4.2.1 Font and emotion

Optical character recognition (OCR) of printed documents has been studied for a long time since the beginning of the computer vision community. As shapes of fonts have become more diversified, including handwritten types, visual font recognition (VFR) has begun to make a mark [174]. VFR has also been applied to OCR for better character recognition regardless of the font used [91]. Recently, high-cost machine learning algorithms such as convolutional neural networks (CNNs) have been utilized to solve large-scale VFR problems [155]. Meanwhile, studies concerning the function of fonts have been gaining attention [9, 79]. The influence of each font on the readability of texts has been investigated [79]. The visual characteristic of a unique font and its emotional effect also studied [9, 111]. These high-level description-based analyses have been applied to font retrieval to realize a better font experience for users [154]. Although the above described works focused on discovering and exploiting the personality of fonts, there is no consideration of exploring emotional influence of fonts.

4.2.2 Psychological emotion modeling theories

There are generally two emotion classification models: categorical emotion states (CES) and dimensional emotion space (DES). CES was developed to define basic emotions based on the hypothesis that a basic emotion is biologically distinct from others. One of the most frequently exploited definitions was proposed by Ekman [45], who defined six basic emotions: happiness, sadness, fear, anger, surprise, and disgust. DES is based on neurophysiology and defines emotions in two (valence-arousal) or three (pleasure-arousal-dominance) dimensions [104, 121]. Figure 4.2 shows the circumflex model of emotion [121], which associates emotion terms with combinations of different intensities of valence and arousal. Based on these psychological theories, there have been many attempts to analyze emotional signals in visual media by the computer vision-related community [102]. However, studies

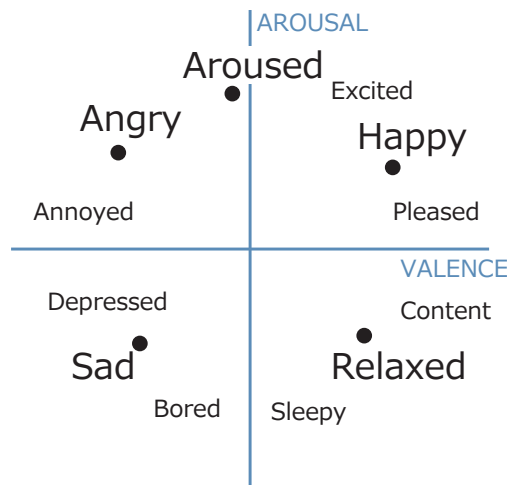


Fig. 4.2 Russell's circumflex model

have never employed emotion classification models to analyze the emotional signals of fonts. In this work, we applied a commonly used psychological model to measuring emotional responses to fonts.

4.2.3 Font as tones of voice

In this study, we see a font as a visual medium that conveys tones of voice. There are studies which share the similar perspective on font. An emotional subtitle system introduced fonts to provide audio cues for hearing-impaired community [112]. This perspective was also utilized in Document-to-audio (DtA) System to render typographic characteristic to auditory modality. Italic and bold fonts are attributed by a small change in the rhythm of speech [47] and changes in pitch and speed of the voice [146] respectively.

Even though there are not many studies that directly map typographic characteristic to auditory modality, we can find indications that relate human auditory system with modulated written text. The effect of manipulated written text on auditory systems has been studied by researchers in cognitive neuroscience, signal processing, and related disciplines [14, 86, 6, 48, 168]. One of the typical method to examined the effect is to observe readers' behavior during silent reading. The indications of *inner voice* experience —the subvocal articulatory rehearsal process during silent reading, have been demonstrated in various ways. Speaker adaptation by exposing to a fast/slow voice sample affected to reading time both silent and oral reading [86, 6]. The eye movement behaviors in silent reading reflected individual's vocal characteristic such as the length of the vowel [48]. It has also been studied that direct speech quotations (e.g., Mary said, "I'm hungry") elicited the more emotional speeches than indirect speech (e.g. Mary said that she was hungry) [168]. There is a study relating the visual

Table 4.1 The label and the number of collected fonts for each category.

Category	Original data (family)	Include font family	Final data
Basic	50	201	28
Fancy	50	72	18
Gothic	50	88	18
Script	50	71	18
Techno& Bitmap	55	83	17
Total	255	515	100

stimuli of written text to acoustic signals of human speech [147]. It modeled emotional responses to bold and italic type for expressive speech synthesis.

Based on the studies, we hypothesize that fonts is a visual cue that demonstrate “how was it said.” To verify our hypothesis, we measure and analysis the unconscious responses to fonts using spontaneous speeches that elicited by fonts of written text in implicit testing.

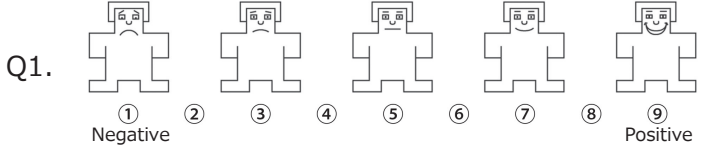
4.3 Font Dataset

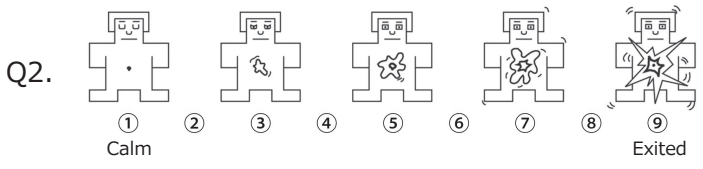
We set up a font dataset for our research. We downloaded a variety of fonts from Dafont¹, which provides access to a huge set of public-domain free fonts. To diversify the appearance of fonts, we collected 50 or 55 font family per category. Dafont has nine categories: Fancy, Foreign look, Techno, Bitmap, Gothic, Basic, Script, Dingbats, and Holiday. The Dingbats and Holiday categories include symbols but not characters, and the Foreign look fonts are hard to recognize as the Roman alphabet. Thus, we excluded these three categories from our consideration. Because the fonts in the Techno and Bitmap categories have similar appearances, we considered them as a single category. Table 4.1 presents the labels of each category and the total number of fonts that we collected. The number in *Original data* column indicates the number of font families. Here, a single font family is consisted of a font and its modifications in stroke width (bold) and orientation (italic), so a single font family contains technically different multiple fonts. The *Include font family* column shows the total number of fonts. In total, we gathered 515 unique fonts.

Because some fonts shared similar visual appearances, we grouped 515 fonts into 100 clusters. We extracted features based on design characteristics [63] and used k-means clustering to divide them into 100 clusters. Because each cluster includes multiple fonts, we determined a representative font for each cluster according to the nearest distance to the cluster centroid. Finally, we built a dataset with 100 fonts.

¹<http://www.dafont.com/>

the quick brown fox jumps over the lazy dog

Q1. 

Q2. 

Q3. 1. Angry 2. Disgust 3. Fear 4. Surprise 5. Happy 6. Sad 7. Neutral

Fig. 4.3 Example user assessment sheet. There are three questions for each font. Q1 and Q2 are for the dimensional emotion space (valence-arousal), and Q3 is for the categorical emotion state (six basic emotions).

4.4 Explicit Testing

In this section, we collect quantitative user labels for each font based on commonly used psychological emotional scale via crowd-sourcing. By analyzing user label data, we will see whether there exists an emotional consensus toward a font. Further, the collected emotional labels will be used to see which fonts has a strong or weak emotional influence on the audience.

4.4.1 Experiment design

Task

Figure 4.3 shows an example user assessment sheet. Participants were requested to fill the Self-Assessment Manikin (SAM) answer sheet, which evaluates the intensity of valence and arousal caused by a given font image (Q1 and Q2 in Figure 4.3). In addition to the DES form, they were also requested to label an emotional term (Q3 in Figure 4.3) by using six basic emotions (happiness, sadness, fear, anger, surprise, and disgust). A task contained 20 assessment sheets corresponding to 20 unique fonts, so we arranged total five tasks for 100 fonts.

Participants

We recruited 72 workers from Yahoo Crowd Sourcing². Each participant was eligible to conduct the task at least once but no more than five times. We paid each worker a reward of \$0.23 to complete a task. We collected 40 assessment sheets per fonts. Because we had 100 unique fonts, we collected 4000 assessment sheets in total.

²<http://req.crowdsourcing.yahoo.co.jp/>

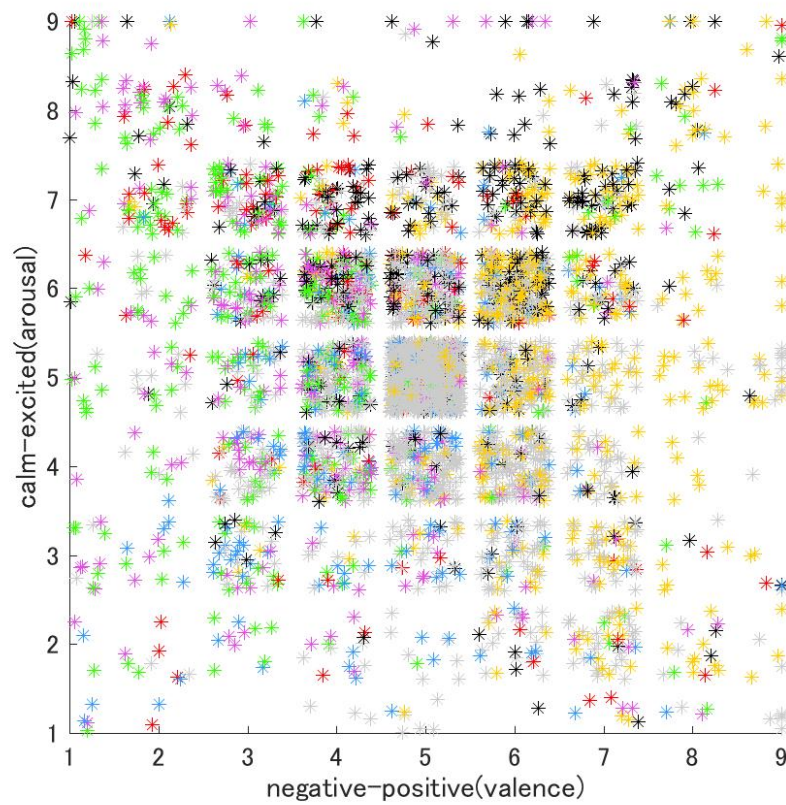


Fig. 4.4 The overall tendency of user assessments. The color of each point indicate the user-labeled emotion category (Angry(red), Disgust(green), Fear(purple), Happiness(yellow), Sadness(blue), Surprise(Black), Neutral(grey)).



Fig. 4.5 Heat map for each emotion to highlight the region where the dimensional assessment was concentrated.

4.4.2 Results

Figure 4.4 shows the overall tendency of 4000 user assessments in the valence-arousal space. The position of each point indicates the valence-arousal level. The color indicates the user-labeled emotion category. We slightly scattered the points by adding random noise within a range of -0.4 to 0.4 to make the results legible. Points with the same color (i.e., labeled in the same category) are located close together.

To visualize Figure 4.4 more clearly, we developed a heat map for each emotion to highlight the region where the dimensional assessment was concentrated (Figure 4.5, the six facial expression images are from here [46]).

4.4.3 Discussion

The heat map results appear to show that there exists a clear valence-arousal difference for each emotion category (Figure 4.5). Also, the distribution tendencies of the front emotions followed Russell's circumflex model well (Figure 4.2). According to Russell, the emotions of anger and fear are distributed in the low valence and high arousal region (quadrant 2), sadness is in the low valence and low arousal region (quadrant 3), happiness is in the high valence and high arousal region (quadrant 1), and surprise (alarm, astonishment) is distributed in the high arousal and middle valence region (Figure

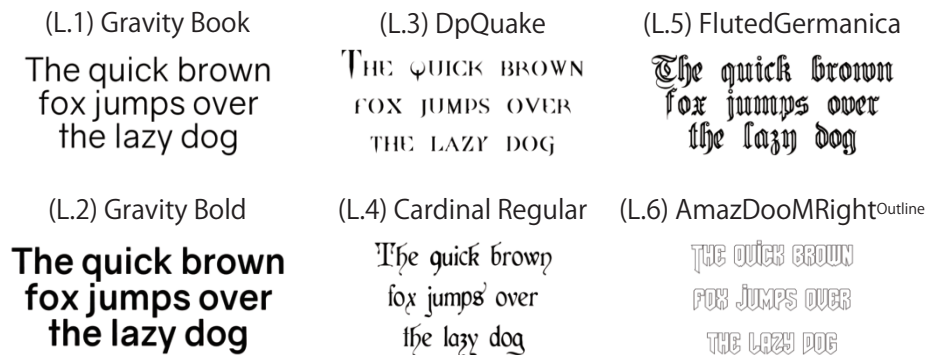


Fig. 4.6 The example fonts with low agreement.

4.2). This implies not only that the quality of assessments from the cloud-sourced workers is reliable but also that the emotions in a font can be analyzed well by a general psychological emotional model.

Disagreement/Agreement in the emotion labeling results

From our crowd-sourced user study, we collected a large amount of emotion-labeled font data. Because the emotion-labeling task is highly subjective, it was difficult to achieve a consensus among workers on a font. Here, we discuss which font had a strong or weak emotional influence on the workers. By analyzing the data, we determined several visual characteristics of fonts that caused disagreement in the emotion labeling results.

Figure 4.6 and Table 4.2 present example fonts with low agreement among users and the assigned categorical ratings. Most general fonts such as normal weighted sans-serif (L.1) and (L.2) had low agreement. We also found that damaged font (L.3) and (L.4), including diminished and stained characters, had low agreement on emotion. Complicated and not very legible fonts (L.5) and (L.6) also had low agreement on emotion due to their complicated and pale appearance. Based on these results, we concluded that general font (L.1 and L.2) and illegible fonts (L.3–L.6) are not appropriate for delivering emotional signals.

Figure 4.7 and Table 4.2 present examples of fonts and their user assessment results with high consensus for each emotional category. There existed design characteristic tendencies for each emotional category. The rounded cap and handwritten style fonts indicated happiness. Slanted and light fonts indicated sadness. fonts with distinctive characteristic received high scores for surprise. Contrast, i.e., variation in thickness between the thickest and thinnest stroke weights, indicated fear. A short and fat appearance and sharp fonts appeared to indicate anger. Slanted and condensed fonts conveyed disgust. We also observed that some emotions were correlated, e.g., fonts that conveyed fear also tended to convey disgust.

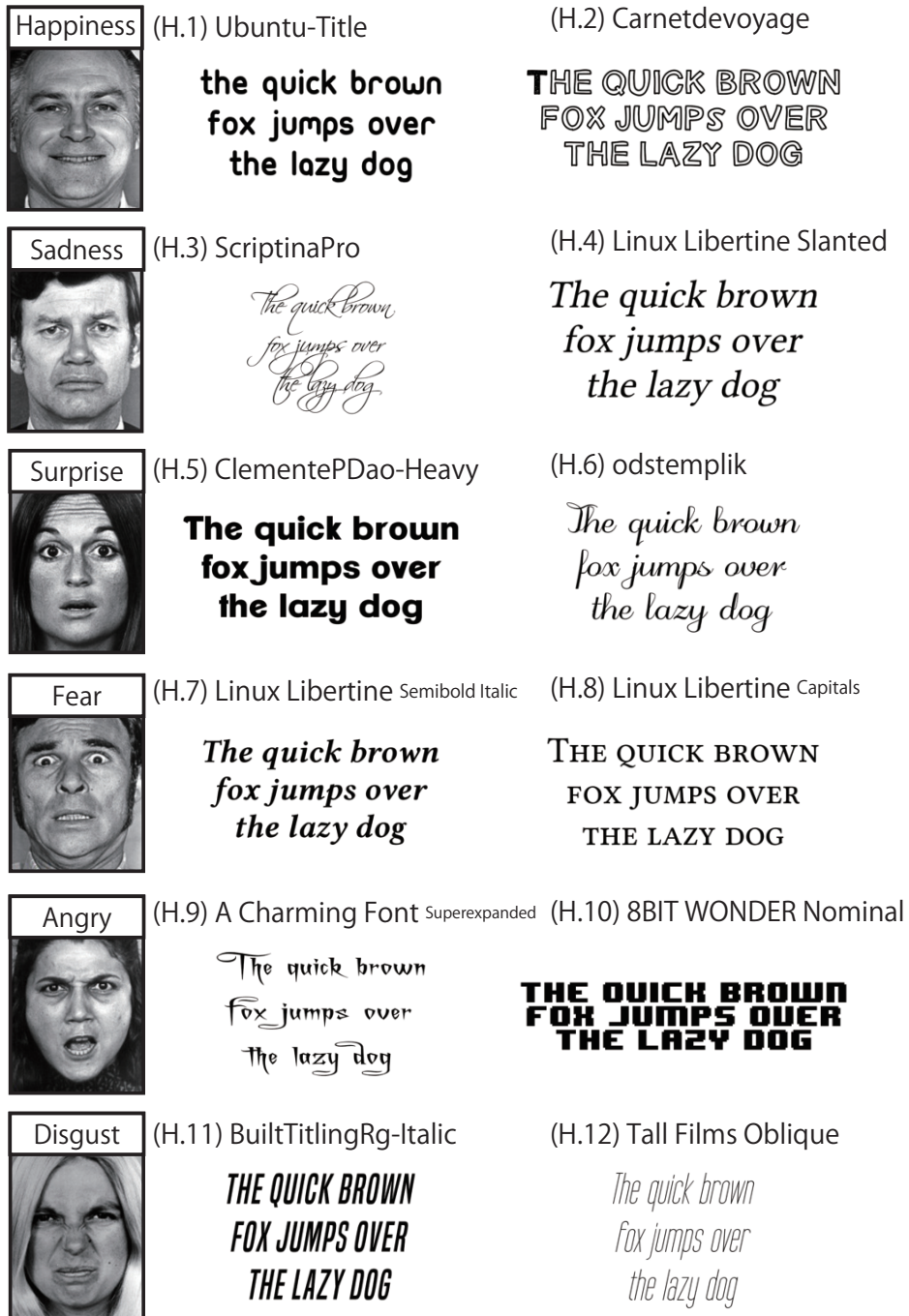


Fig. 4.7 High consensus font examples for each emotional categories

Table 4.2 The user assessments results of high consensus font examples (Figure 4.7) and the user assessments results of low consensus font examples (Figure 4.6). Ha, Sa, Su, Fe, An and Di are abbreviation for Happiness, Sadness, Surprise, Fear, Angry and Disgust respectively. Here, we exclude the number of votes for neutral.

High consensus	Ha	Sa	Su	Fe	An	Di
(H.1)	9	1	0	0	0	0
(H.2)	10	3	0	0	0	0
(H.3)	4	12	3	1	2	1
(H.4)	2	15	0	7	0	6
(H.5)	9	1	20	0	1	2
(H.6)	6	1	15	0	0	1
(H.7)	0	2	2	15	0	19
(H.8)	1	4	0	17	5	12
(H.9)	3	2	7	0	11	7
(H.10)	3	2	7	5	11	3
(H.11)	1	1	2	6	2	21
(H.12)	1	4	4	3	1	10
Low consensus	Ha	Sa	Su	Fe	An	Di
(L.1)	4	4	4	0	0	0
(L.2)	11	11	1	0	0	1
(L.3)	4	4	4	5	4	6
(L.4)	4	8	3	1	0	6
(L.5)	6	4	4	2	1	1
(L.6)	6	0	5	3	3	1

4.5 Implicit Testing

In this section, we measure unconscious response to fonts. To do that, we examine the effect of fonts on oral reading by providing prosodic analysis for speeches obtained from different font stimulus. In addition to the analysis, we introduce the subjective evaluation on the obtained speeches to determine how well people recognize the emotion naturally induced by the font.

4.5.1 Experiment design

Reading material

We prepared three positive stories and three negative stories referring to the reading materials provided in [167]. Table 4.3 shows an example of negative story. As we can see, each story is composed of three paragraphs, i.e., background, focus and target (see Table 4.3). After giving reading material, we provided comprehension test to make participants to engage in reading task. We manipulated the target paragraph in three conditions: 1) normal; 2) font; and 3) instructed. The text in the normal condition is written in the neutral font. The target text for positive or negative story in the font

Stimuli	1) Normal	2) Font modulated (Font group)	3) Instruction modulated (Instructed group)
Background sentence	A trendy night club had been recently opened near Kerry's flat. Every night, Kerry was terribly disturbed by the thudding noise.		
Focus sentence	Kerry said		
Target sentence	It is so ridiculous every time	<i>It is so ridiculous every time</i>	(In angry voice) It is so ridiculous every time
Comprehension test	Q. Did Kerry feel negative? - Yes - No - Neither		

Table 4.3 An example of negative story. It is composed of three paragraphs, i.e., background, focus and target. After giving reading material, we provided comprehension test to make participants to engage in reading task.

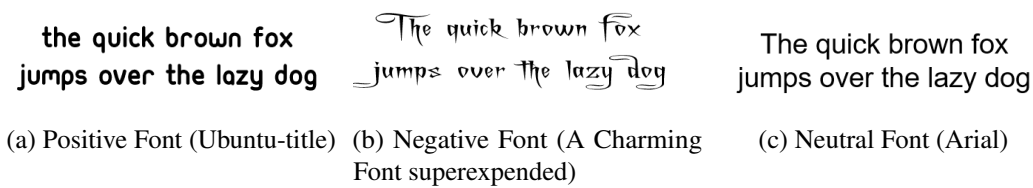


Fig. 4.8 Three fonts that we used in our implicit testing.

condition is written in the positive or negative font respectively. In the instructed condition, we gave the text written in the neutral font, but we guided them to read text "in happy/angry voice."

Selection of neutral, negative and positive fonts

In Section 4.3, we collected font dataset and corresponding emotion labels. Considering the legibility of fonts, we picked two fonts that had the highest positive score and negative score from the dataset (Figure 4.8a and 4.8b). As we discussed in the Section 4.4, we selected the font *Arial* as default font that has low emotional influence (Figure 4.8c).

Participants

There are two groups of participants, i.e., font modulated and instruction modulated groups. The font modulated group read the given reading materials under two conditions, *normal* and *font*. The instructed group read the given reading materials under two conditions, *normal* and *instruction*. We recruited eight participants (4 male) for the font stimuli group and four participants (2 female) for the instructed group. Informed consent was obtained under an understanding of information

about the experiment, including voice recording process. No participants had knowledge for the experiment. Most of them are non-native English speakers, who are fluent bilingual speakers. Because the emotional prosody features we focused are independent on speakers' mother tongue (but less expressive) [131, 4], we did not consider the mother tongue as a variable. In other words, participants in a group (font group) read materials under implicit emotion modulator, i.e., font, and participants in another group (instructed group) read materials under specified instructions.

Environment setting and tools

Participants sat about 20-30 cm in front of a 14.0" (1366x768, resolution 720p) Dell laptop with integrated microphones, and each speech was recorded with 44.1kHz sampling rate and two-channel 16-bit sampling format. Participants in the font group generated a pair of two speeches under two conditions, i.e., normal and font for a given story. This resulted in total 96 speeches: six stories \times two conditions \times eight participants (a pair missing, total 94 speech samples). Likewise, participants in the instructed group generated a pair of two speeches under two conditions, i.e., normal and instruction for a given story. This resulted in total 48 speeches: six stories \times two conditions \times four participants. After denoising the recorded speeches with audio editing software Audacity, we segmented and labeled the speeches with ProsodyPro, a Praat script [161] in two ways — paragraph level and word level (see Figure 4.9). The stories were segmented into three parts: background, focus and target (paragraph level). Then the target paragraph was further divided into word (word level). Using Prosody Pro, we could get average pitch, duration, excursion size and intensity for each labeled segment from the collected speeches (Figure 4.9).

4.5.2 Paragraph level analysis

Analysis method

Even if the same speaker uttered the same story under the same condition, the prosodic features vary on each try [66]. To avoid this biased effect, we defined a relative score, which refers how the prosodic feature of the target utterance has changed compared to the previous utterance in the same story reading session. Figure 4.9 illustrate the analysis method we designed. For example, Relative Pitch Change (RPC) measures the increment in pitch from the focus paragraph to target paragraph, and we formularized this as:

$$RPC(focus, target) = \frac{f0_{target} - f0_{focus}}{f0_{target} + f0_{focus}} \quad (4.1)$$

Here, $f0_{focus}$ and $f0_{target}$ refer the average pitch over the focus paragraph and target paragraph respectively. The Relative Intensity Change (RIC), Relative Excursion Change (REC) and Relative Duration Change (RDC) can be calculated in the same manner. Then we compared two speeches generated by a single speaker under two different conditions, e.g., font and normal. For example, if a speaker i generated a higher pitch speech under the font condition than that of the normal condition

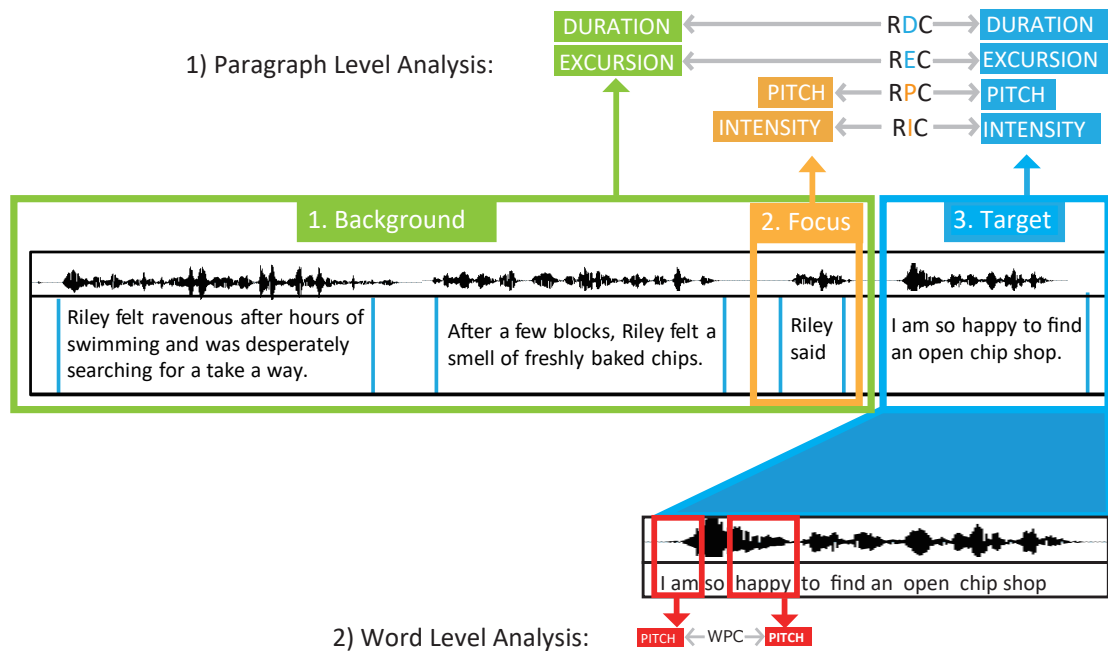


Fig. 4.9 An example of a positive story and two analysis method (paragraph level and word level).

Score of positive stories	Font	Instructed
Excursion (REC)	0.792*	0.917*
Duration (RDC)	0.696*	0.667*
Pitch (RPC)	0.696*	0.917*
Intensity (RIC)	0.391	0.889*
Word Pitch (WPC)	0.739*	0.667*
Score of negative stories	Font	Instructed
Excursion (REC)	0.417	0.667*
Duration (RDC)	0.833*	0.917*
Pitch (RPC)	0.588*	0.750*
Intensity (RIC)	0.500	0.889*
Word Pitch (WPC)	0.708*	0.750*

Table 4.4 The result of our prosodic feature analysis. The higher score indicates that a certain prosodic feature (REC, RDC, RPC, RIC, or WPC) was more pronounced in the modulated by Font stimuli or Instructed stimuli than that of Normal. We can see that implicit responses by font stimuli seem to follow the tendency of explicit responses by instructed stimuli (* : significance@0.05 level in Binomial test).

for a given story j , $Val_{i,j}$ is 1, otherwise 0.

$$Val_{i,j} = \begin{cases} 1, & \text{if } RPC_{font} > RPC_{normal} \\ 0, & \text{otherwise.} \end{cases} \quad (4.2)$$

The score indicates the relative pitch increase effected by font and it is calculated as follows:

$$Score_{RPC}(font, normal) = \sum_i^P \sum_j^S \frac{Val_{i,j}}{P \times S}, \quad (4.3)$$

where P and S indicate the number of participants and stories, respectively. The higher the score is, the more font effect is. Not only for the font group, we also calculated the score of instructed group (e.g., $Score_{RPC}(instructed, normal)$), and same process has been conducted on the other prosody features too, i.e., RIC, REC, and RDC.

Results

Because the acoustic characteristics vary according to each valence of stories (positive/negative), we obtained scores separately on the valence of the story. Table 6.2 shows the calculated scores by the valence of the story, and by the group. If the prosodic feature of the target paragraph in the font (or instructed) condition is greater than that of the normal condition, the score of the prosodic feature reaches 1. On the other hand, if there is no difference in speeches between the normal condition and the font condition (or instructed), the score reaches 0.5 (random level). From the table, the speeches

collected in the instructed condition are characterized by broader excursion range, longer duration, higher pitch and higher intensity. We conducted binomial tests to examine the difference is not due to chance. We regarded the results of the instructed condition as the ground truth, which prefigures how the results in the font condition should be.

According to our analysis, speeches collected in the font condition are described by longer duration, and higher pitch in both positive and negative stories, and broader excursion range in the positive story task. From this result, we can say that font contributed to generating emotional speeches which have similar prosodic characteristics to instructed speeches. In the case of excursion, however, the score of the negative story was not marked or rather lower than the speeches in font condition (0.417). This can be explained as follows. The prosodic features of anger are determined by many factors. One of them is from the difference in expressing anger by individuals, i.e., hot-anger or cold-anger. In general, acted emotions by instruction tend to be more expressive [157, 100], and are more likely to have higher arousal states which characterized as the broader excursion range. We confirmed this tendency in our negative story results showing the broader excursion range in the instructed condition (Table 6.2). However, compare to the results in the instructed condition, we can see that the font did not contribute to the excursion range of the speeches. Here, we can surmise that the font condition made participants to express emotions on their voice less consciously than the instructed condition, and this resulted in different effects on excursion range across all the reading tries.

As similar to the excursion range tendency in the negative story task, there were not significant contributions to the intensity of speeches attribute to font usage. In other words, both positive and negative speeches did not follow the intensity characteristics in instructed condition. Given that intensity is an indicator of highly aroused and intended emotions [158, 149], it appears the emotional speeches in the font condition were not generated with a strong intention. In terms of pitch, though we can say that the result in the font condition with the negative story task shows a significant difference, the result in the font is not clear as not much as the results in the instructed condition. This will be discussed further in the word label analysis (Section 4.5.3).

4.5.3 Word level analysis


Analysis method


People usually emphasize a sound on words that describe emotions which are important to the meaning of the sentence. In the word level analysis, we investigate whether the font contributes to greater pitch emphasis to these keywords for expressing emotions. We defined Word Pitch Change (WPC), which refers the pitch increase from the first word of the sentence (e.g. *I'm*) to the emotional keyword (e.g. *happy*) as follow (Figure 4.9):

$$WPC(I'm, happy) = \frac{f0_{happy} - f0_{I'm}}{f0_{happy} + f0_{I'm}} \quad (4.4)$$

Rating 1

1) Choose the speech clip that sounds more emotional.

Speech A


Speech B


2) Choose the sentence which matches to what you've just heard among the given choices.

I'm happy to drive you again

What a rip-off flight charging for fee

I really missed you so much

It is so ridiculous every time

This is so relaxing after a busy time

I really hate the winter in Japan

It's about to start

I am so happy to find an open chip shop

Fig. 4.10 An example of the assessment sheet

Then we calculated the score which indicates the font effect on emotional keywords in the same manner in Chapter 4.5.2 (see Equation 4.2 and 4.3).

Results and discussions

Row Word Pitch Changes (WPC) in Table 6.2 shows the calculated scores. The higher WPC score is the more font effect. In both instructed and font conditions, participants generated speeches putting more emphasis on the sound of emotional keywords than that of the normal condition.

In the paragraph level analysis (Chapter 4.5.2), we observed participants generated speeches showing higher in pitch with the negative stories, but the difference was not much clear as the positive story result. Considering both the paragraph level and word level results, the participants in the font condition seemed to generate more emotional speeches, and this observed in more emphasis on the emotional keywords and higher average pitch in the font condition than that of the normal condition.

4.5.4 Subjective evaluation on the collected speeches

In addition to the prosodic characteristic analysis, we conducted a subjective evaluation on the collected speeches to investigate whether people recognize emotions affected by font stimuli or not. We made use of Amazon Mechanical Turk (MTurk) service. MTurk workers conducted a

Speakers	Positive	Negative	Total
1	0.672*(134)	0.478 (134)	0.575* (268)
2	0.716*(134)	0.594*(133)	0.655* (267)
3	0.733*(135)	0.583*(132)	0.659* (267)
4	0.542 (131)	0.386 (132)	0.464 (263)
5	0.411 (129)	0.674*(89)	0.518 (218)
6	0.718*(131)	0.704*(135)	0.711* (266)
7	0.731*(134)	0.534(131)	0.634* (265)
8	0.681*(135)	0.621*(132)	0.652* (267)
Total	0.652*(1063)	0.568***(1018)	0.611*(2081)

Table 4.5 The subjective evaluation from the crowd-sourced study. The number in brackets refers to the number of assessments. (* : significant@0.05 level in Binomial test)

two-alternative forced choice task consisting of two options, one of which is the speech obtained in the font condition, and another is the speech obtained in the normal condition. In each task, workers were forced to choose one of these options which sounds more emotional.

Experiment design

Ninety-nine Mechanical Turk workers evaluated 47 pairs of two speeches from the two different conditions. Figure 4.10 shows an rating example. An assessment includes two questions, one question for evaluation and one for screening malicious workers. Each worker listened the two speeches and chose the speech that sounds more emotional. We allowed workers to listen the speech clips repeatedly as they want. The screening question asked workers to choose the sentence they heard in the same assessment. A HIT, which refers a smallest unit of task consists of twelve assessments, and we paid \$0.1 for completing a HIT. We controlled the order effects with counterbalanced task design.

We obtained fifty HITs per speech pair that results in total 2,350 assessments. We reject HITs which marked the answer sheet randomly (e.g. AAAA...AABBAABB...), and completed in 2 min and assessments which failed to correctly answer to the screening question. Consequently, we collected 2,081 assessments.

Results

Table 4.5 shows the result of the subjective evaluations from the crowd-sourced study. We conducted binomial tests to examine the difference is not due to chance. Asterix mark indicates that the difference is significant at 0.05 level in Binomial test.

The value of each cell indicates the proportion of workers who answered that the speech obtained in the font condition sounds more emotional. Each row indicates the results from each speaker's voice samples. The number in brackets refers to the number of total assessments obtained by MTurk workers. The final row is the average results of all speakers.

As we can see in the Table 4.5, the font effect varied with the individual who generated speeches. In case of the speaker 4, we cannot observe any font effect. Actually, the speaker produced a very low tone of voice with the frequency between 80Hz-90Hz which made it difficult to recognize emotions by the workers. Nevertheless, we observed a general tendency that the speeches in the font condition recognized as more emotional, and the results of positive speeches reflect the stronger font effect than the negative speeches.

4.6 Conclusions

In section 4.4, We showed that existing emotional models can be applied to font emotion analysis problems through a crowd-sourced user study. We observed not only that fonts convey unique emotional signals but also that there existed strong and weak emotional consensuses towards fonts among users.

In section 4.5, we investigated the simultaneous effect of the visual characteristic of font on the oral reading. The prosodic analysis result showed that the speeches obtained under font condition followed the characteristics of the speech samples that were obtained when we asked to act specific emotions —longer duration, higher pitch, and broader excursion range. In addition to that, MTurk workers reported that the speeches obtained in the font condition sounded more emotional than the speeches obtained in the normal condition.

The experiment results give persuasive evidence that the unusual font simultaneously affects the reader's emotion, and makes the reader to unconsciously produce more expressive speech than the commonly used font text. The findings of this study could provide a foundation for suggesting promising applications for speech synthesis, vivid reading experience of fairy tales, etc.

Chapter 5

Emotype: Expressing Emotions by Changing Typeface in Mobile Messenger Texting

Instant messaging is a popular form of text-based communication. However, text-based messaging lacks the ability to communicate nonverbal information such as that conveyed through facial expressions and voice tones, although a multitude of emotions may underlie the text of a conversation between participants. In this paper, we propose an approach that uses typefaces to communicate emotions. We investigated which typefaces are useful for delivering emotions and introduced these typefaces into a mobile chat app. We conducted a survey to demonstrate how changes in the typeface of a message affected the meaning of the message conveyed. Our user study provides an understanding of the actual user experience with the application. The results show that the use of multiple typefaces in a message can affect and intensify the valence received by users and the use of multiple typefaces elicited an active response and brought about a livelier mood during texting.

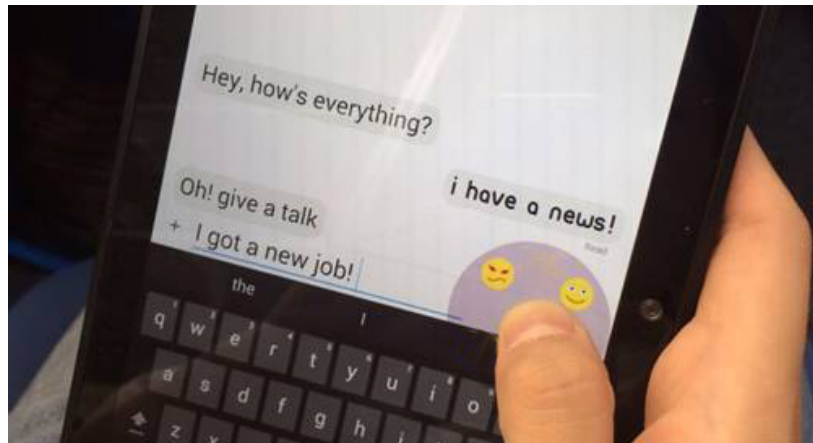


Fig. 5.1 User interface of the Emotype prototype. A user can send a text message and change the message typeface to convey a particular feeling.

5.1 Introduction

The sharing of emotions among people can elicit empathy, increase a friendly feeling, and even improve mental health [35]. In face-to-face situations, people can express their emotions or feelings with facial expressions, voice tones, gestures, and so on. The function of emotional expressions is not only to express one's inner state, but also to guide a situation toward an intended mood [41]. For example, if someone asks, "Could you do me a favor?" with a smile, the mood of the situation may be positive, and the receiver of the question would likely be more willing to respond affirmatively. However, if the person makes the request in an overbearing tone of voice, the mood may be negative, and the receiver may feel coerced to respond. In this way, nonverbal signals can change the receiver's attitude, which can affect his or her reaction to the situation. Despite the importance of nonverbal cues in effective communication, computer-mediated communication tools such as text-based communication (e-mail or text messaging apps) inherently lack nonverbal signals for conveying emotional information [150].

As the text-based communication become popular, alternate ways for expressing nonverbal signals such as emoji have developed. These pictorial representations are contributing to our communication engagement by making the communication more lively. Meanwhile, the effect of typeface has been studied in the design and marketing literature such as emotional response to logos and advertisements [63, 96], preference for use of certain fonts [135], and the relationship between the visual appearance and emotional responses in fonts [9]. However, the role of typeface has been limited to marketing, so that only experts, such as designers, have considered the use of typefaces as a communication medium.

In this paper, we propose Emotype, a mobile messenger application prototype that enables general users to change the typeface of a mobile messenger message to convey certain emotions (Figure 5.1). The novelty of this study is to generate a new value of typeface in the context of the mobile messaging,

where the role of typeface has not been investigated. Even though there are already various ways of expressing emotions in mobile messengers (emoticon/emoji, voice message, and so on), introducing a new system using typeface will bring a richer communication experience. We expect our prototype enables not only communication experts but also general users to become aware of the emotional function of the typeface so that they actively engage in font communication.

The contributions of this study are outlined below. We:

- build a mobile messenger application prototype that enables users to communicate with typefaces;
- demonstrate the feasibility of typefaces for communicating emotions with a survey study; and
- explore the unique feature of typeface different from other ways for expressing emotion by qualitative user study.

As far as we know, this paper is the first work to explore the effect of typefaces for message texting. We therefore focus on the basic emotion scale (negative, positive and neutral [55]) to observe the sharp contrast of effects between negative and positive fonts. In the following section, we review related studies about typefaces in communication. We then present the Emotype application prototype and introduce the emotion-bearing typefaces adopted in the prototype. Our user study design is then introduced. We investigate the effect of the message content in combination with positive and negative typefaces and present various and unique user behaviors, experiences, and user comments obtained from the user study.

5.2 Related Works

5.2.1 Nonverbal signals in mobile messengers

Users of mobile messengers cannot express their emotions through the device using facial expressions, voice tones, or gestures. Nonetheless, alternate ways exist for expressing nonverbal signals in mobile environments. Emoticons, which are a combination of punctuation marks and letters that represent human facial expressions, are one of the most general ways to express nonverbal signals by mobile messenger users (e.g., “:-)”, smiley face). In the late 1990s, emoji (e.g., smiley) were invented and added to the Unicode system and became increasingly popular with the development of devices supporting graphics, such as smartphones. Emoticons and emoji perform the role of providing emotional signals, thereby improving communication, and conferring a certain mood to the chat [116]. Capitalization of all letters can provoke polarized responses. For example, the use of full capitalization of a positive sentence, such as “HAPPY TO HEAR THAT,” conveys extreme joy in contrast to the typical use of upper and lower cases, such as “Happy to hear that.” In the same way, capitalization of a negative sentence intensifies the negative feelings [20]. It has also been verified that letter repetition, which extends a syllable, such as “Sweeeet”, conveys auditory signals and invokes

playful impressions [80]. Moreover, punctuation and quotation marks can be used to convey sarcasm [40], such as “Thank you ‘very’ much.” It is also known that using more words or being quick to respond can engender a positive impression in the receiver [59]. In addition, the size of the text affects the sense conveyed. For example, the Google Allo messaging app enables users to change the size of the font to denote voice volume [74]. Recently, haptic interface demonstrated that emotions can be transferred by tactile stimuli such as temperature [145].

5.2.2 The usage of typeface in multimedia

The importance of the typeface has been investigated in many fields, including design, marketing, and computer vision. The legibility of a typeface is important for effective communication and has thus been studied for many years [128][12][7]. It has been shown that the arrangement of typefaces, e.g., kerning, line spacing, and letter spacing, affect communication. Font size and line spacing are classic issues that affect legibility and comprehensibility of text [127]. In addition to the function of a typeface in effective delivery of content, the impressions created by a typeface have been actively discussed in recent years. Many researchers have suggested that each typeface creates a unique impression, which is also known as the typeface’s personality [135][9][96]. Shaikh et al. investigated the relationships between personality traits and preference for use of certain fonts [135]. They suggested that sans serif fonts are effective for website text or email, but serif fonts are more effective for business documents. Li and Suen examined the personalities of 24 typefaces and grouped them into four categories—directness, gentleness, cheerfulness, and fearfulness—based on survey data [96]. Amare and Manning explained why specific typeface features elicit certain emotional responses, thereby supporting the findings of previous studies based on empirical user evaluations [9]. These typeface personalities have been studied and utilized in commercial applications, such as advertising and market research.

As the number of typeface increasing, font retrieval challenge has gained attention. Thanks to the recent advances in computer vision, font identification from real-world text images such as signboard has achieved remarkable performance for both English [154], and Chinese fonts [70]. These systems are helpful when a user has a certain reference image for the typeface to be searched. Search systems that can be applied in another search scenario have also been proposed; exploring fonts using high-level attributes, such as “dramatic” [111] and recommending fonts that match well with users’ graphical input such as background image [31]. Recently, by leveraging recent generative adversarial networks (GAN), typeface glyph synthesizing from few samples by separating style from text image has developed [13, 172]. The above studies support the idea that each typeface has its unique style and affects the sentiment of the written text. Differing from those works, we explore emotional effects of typefaces for message texting.

5.2.3 Typography communication in mobile messengers

Several mobile applications exist that enable users to exploit the emotional effects of typefaces. The Font Dresser [106] font editor application, for example, enables users to change the typeface, text color, and so on. However, it requires saving the text of the changed font into an image format to share it. The Font Infinity [98] application employs symbols that look like roman characters to create fancy letters (e.g., H∞ϑ ι† ℓ∞κs?). However, the function of these apps is limited to an aesthetic role. Recently, stickers and illustrations, such as emojis, on Line [34] mobile chat apps provide text with various typefaces in the form of illustrations. The use of OCR font in web-based advising service made users perceive a adviser as a chatbot [22]. Google's Allo is a chat application that allows users to change the size of fonts. Text in a small font size implies a whispering voice, while text in a large font size conveys a raised or shouting voice [74]. Kinetic typography — moving text, emotionally influences the viewer through changes in animation, speed, and dynamics of written text [94][82].

These applications reflect the need for general users to exploit the various effects of typefaces. However, compared with emoji and emoticons, the affect associated with typefaces on the mobile environment has not been studied. Although [82] studied the effects of kinetic typography, there have been few presentations of actual user experiences.

5.2.4 Studies on affect and emotion

Two frequently used psychological emotion models exist: dimensional and categorical. The dimensional model defines emotions in a continuous space in two or three dimensions (valence, arousal or include intensity) [104]. In the categorical model, emotions can be described as discrete categories, e.g., six emotion categories [45] (happiness, sadness, anger, disgust, and fear) and the polarity scale [55] (negative, positive and neutral). These psychological models have been applied not only to recognize emotions from various media, including images, text, and audio [102][83][8], but also to collect typeface emotion labels [33].

Many works have studied about various personality traits (or sentiments) of typeface. Some researchers investigated emotions created by typefaces (e.g., happy, angry) based on the general emotional models stated above [114], while others considered impressions [96] (e.g., friendly, reliable). Other studies have examined the function of typefaces for delivering content [128][29] (e.g., readability and memorability).

In this paper, we investigate whether typeface contributes to change the state of valence and intensifies the emotion a message conveys. As dimensional model theory (valence-arousal) explains, intensified emotions are likely to have a high arousal showing u-shaped curve in the valence-arousal dimension [107]. We, therefore, focus on the polarity of typeface — negative, positive, and neutral, instead of using the models which classify emotions sensitively by the level of arousal.

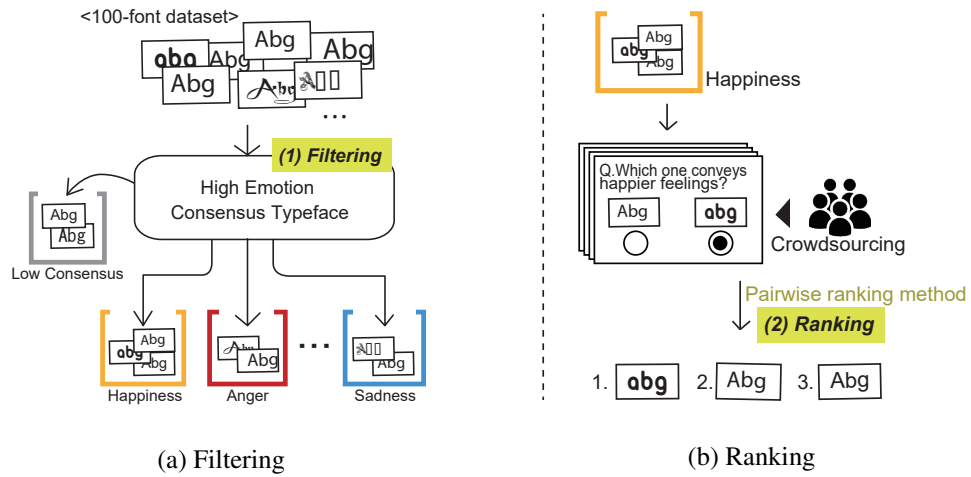


Fig. 5.2 Example of a two-alternative forced choice (2AFC) question.

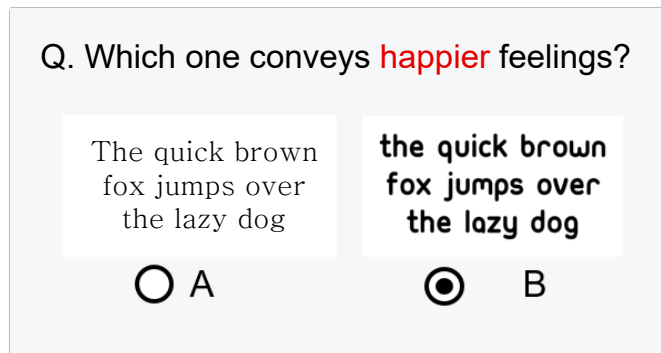


Fig. 5.3 Example of a two-alternative forced choice (2AFC) question.

Positive	Neutral	Negative
handgloves	Handgloves	HANDELLOVES

Fig. 5.4 Representative typefaces for each emotion category.

5.3 Material

5.3.1 Emotional typefaces

Choi et al. developed 100-Font dataset [33] in which each typeface was labeled with six emotion categories (*happiness, sadness, surprise, fear, anger, and disgust*) by 40 crowdsourcing workers. As we discussed in the related work, this work focuses on the polarity of typeface. Therefore, given the dataset with six emotion categories, we selected two representative emotions for the polarity scale — happiness and anger which are both high arousal emotion. By doing that, we expect to be able to observe the sharp contrast between positive typeface and negative typeface which illustrates effects of the typefaces.

Because the dataset only provided the overall labeling tendency which shows that some typefaces have a high agreement in the labeling result among workers, we investigate to find the fonts which have the most emotional influence in each emotion category. Figure 5.2 shows the procedure how we found the most emotional font for each emotion category. Because some of the fonts have a low agreement in the labeling results, we firstly filtered out typefaces that have a low emotional consensus (Figure 5.2a). Then, in order to select the most emotional font for each emotion among the filtered fonts, a crowdsourcing study was designed (Figure 5.2b). We recruited 200 workers from Yahoo crowd sourcing service, and they conducted the two-alternative forced choice (2AFC) task (Figure 5.3). Based on the collected workers' assessments, the emotional influence scores were estimated using a pairwise ranking algorithm [27]. Here, we chose anger font among the negative fonts and regarded it as a negative font for the following user studies. In the same way, we regarded the happiness font as a positive font. And the font *Arial* is selected as a neutral font.

Figure 5.4 shows the representative typefaces that have the highest emotional influence score for two categories via pairwise ranking study. The positive (happiness) font has rounded caps and a well-balanced appearance. The negative (anger) font has heavily weighted characters on a wild painted background.

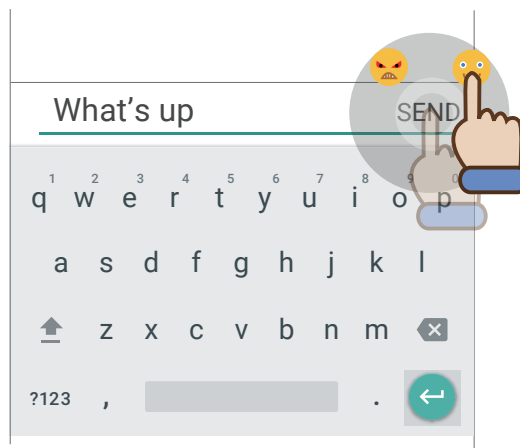


Fig. 5.5 An envisioned interface for changing the typeface.

5.3.2 Messenger application prototype

We developed a messenger application that enables users to easily change the typeface of a message to reflect their emotion. Figure 5.5 shows the illustration of the Emotype prototype interface. Emotype was developed based on the Atlas [93] open-source customizable messaging app.

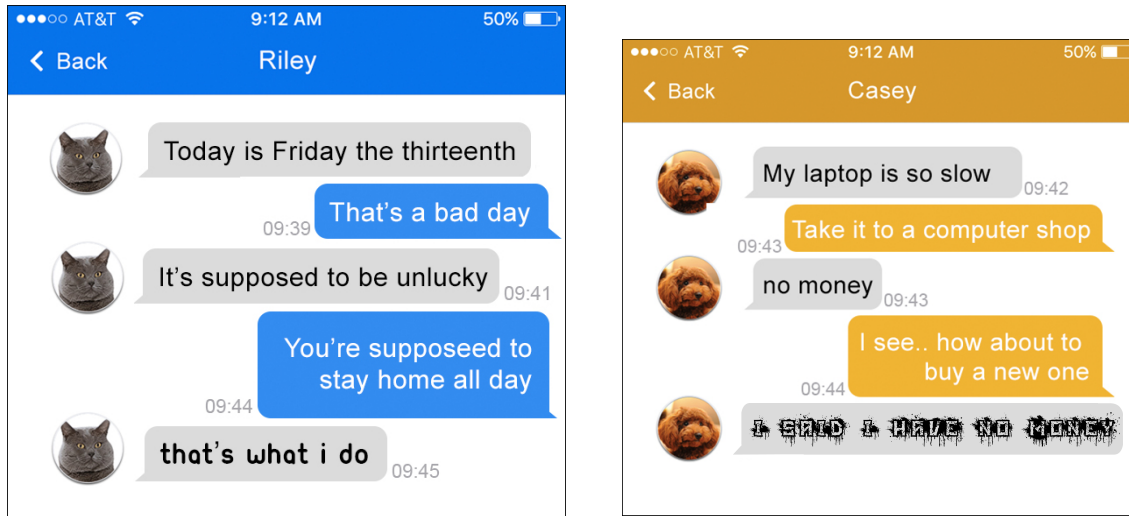
The Emotype mobile application enables users to select a typeface when they want to send a typed message (Figure 5.5). When a user applies a long touch on the *SEND* button, the emotion selection window appears. The user can then move his or her finger to the intended emotion. With these gestures, the typeface of the typed message changes from the neutral to the emotional typeface that the user intended. If the user simply tap the button *SEND* without the swiping gesture, the message is sent in the *Neutral* typeface.

5.4 Quantitative Study: Feasibility of Using Typefaces for Emotion Communication

To demonstrate the feasibility of using typefaces for emotion communication, we tested two null hypotheses:

- H_{01} : the typeface used in a message cannot affect the valence (positive or negative) received by users;
- H_{02} : the typeface used in an emotive sentence cannot intensify the valence (positive or negative) received by users.

We tested the hypotheses using a questionnaire. We created an online survey that was compatible with both mobile devices and desktop computers. We invited 55 participants who had no background



(a) H_{01} : Testing whether typefaces affect the valence of a message (Task Neutral 2 in Table 5.1).

(b) H_{02} : Testing whether typefaces intensify the valence of a message (Task Negative 1 in Table 5.1).

Fig. 5.6 Examples of conversations for testing each hypothesis.

information about the study, and 36 participants (20 males, 16 females) between the ages of 17 and 46 years old (mean = 28.42, SD = 5.07) completed the questionnaire fully.

5.4.1 Task

We provided various conversations to participants (Table 5.1). Figure 5.6a and Figure 5.6b show examples of the tasks for testing H_{01} and H_{02} , respectively. The lines in bold for each conversation indicate the target messages that have variations in the typeface used. Participants were requested to imagine the speaker's facial expression and voice tone during the target message.

Task 1: Questions for H_{01}

We demonstrated that the valence on a message is changed by the typeface used. To do that, we designed the target message to be ambivalent (refer to rows in H_{01} of Table 5.1). Participants were requested to guess the speaker's facial expressions and voice tone during the target message and to pick the option that gave the closest match (−1: negative, 0: neutral, 1: positive). We prepared three conversations, and each conversation had three variations in the typeface used (negative, neutral, and positive). In total, participants responded to nine questions that were given in random order.

Task 2: Questions for H_{02}

We investigated whether the perceived emotional effect was strengthened by the positive or negative typeface. For that reason, we designed the target message to be negative or positive (refer to rows in H_{02} of Table 5.1). We asked participants to guess how much the speaker expresses negative or

Table 5.1 The conversations provided for testing the two hypotheses. The lines in bold for each conversation indicate the target messages that have variations in the typeface used. For $H0_1$, we designed the content of the target message to be neutral with three typeface variations (neutral, positive, and negative). For $H0_2$, the target message was obviously negative or positive, and each has two typeface variations (neutral and negative vs. neutral and positive).

	Task	Conversation
$H0_1$	Neutral 1	A: I visited an art exhibition B: Anything interesting? A: There was a painting of a jar that was full of pencils. But the artist said the jar was both full and empty B: but it was full of pencils! how could he say it was empty? A: Artists see things differently
	Neutral 2	A: Today is Friday the 13th B: That's a bad day A: It's supposed to be unlucky B: You're supposed to stay home all day A: That's what I do
	Neutral 3	B: Where are you gonna go? A: I have to walk the dog B: What kind of dog do you have? A: poodle B: Oh, they bark a lot A: They sure do
$H0_2$	Negative 1	A: My laptop is so slow B: Take it to a computer shop A: no money B: I see. How about buying a new one A: I said I have no money
	Negative 2	A: Did you hear about the baseball player? B: The home run hitter on drugs? A: Yeah. I'm a big fan of that team... B: It caused a \$7 million loss to the team A: I want to beat him up
	Positive 1	B: I am eating a simple salad A: What do you put in it? B: Just lettuce, tomato, and celery A: That's it? B: Then, I add French dressing A: That sounds good
	Positive 2	B: Gravity is very important A: What is gravity? B: the force that pulls everything down A: I don't understand B: You would float into the sky like a balloon A: That would be fun

Table 5.2 The values in each cell indicate MEAN (SD). The target messages with positive or negative typefaces achieved higher positive or negative ratings. Here, a higher score means more positive, a lower score means more negative. The values with an ** mark indicate a statistically significant difference at the .01 level.

	Neutral	Positive	Negative
Neutral 1**	0.08 (.43)	0.36 (.58)	-0.50 (.69)
Neutral 2**	-0.06 (.47)	0.17 (.65)	-0.53 (.73)
Neutral 3	0.03 (.44)	0.03 (.55)	-0.69 (.57)

positive feelings in the last message and they were requested to respond using a five-point Likert scale (1: not much, 2: a little, 3: somewhat, 4: much, 5: a great deal). We prepared two conversations for each emotion (Positive 1 and 2 for positive, Negative 1 and 2 for negative) and each conversation had two variations in using typeface (neutral or emotional). Consequently, there were eight questions and they were given in random order.

5.4.2 Results

Table 5.2 shows the mean and standard deviation of the valence ratings by typeface to test H_{01} . For Neutral 1 and 2, the target message with positive typeface brought about higher scores than others (higher scores indicate more positive). In case of the message with negative typeface, participants rated lower scores than others (lower scores indicate more negative). For Neutral 3, we could not discover any mean difference between neutral and positive typefaces. However, the target message with negative typeface brought about much lower ratings than those with the neutral typeface. To assess the significance of differences between means for each typeface, one-way ANOVA was conducted (for conversations 1 and 2 at the .01 level). We found a significant effect of typeface on the perceived valence (Neutral 1: $F(2,105)=20.25$ $p < .001$, Neutral 2: $F(2,105)=11.36$ $p < .001$).

Table 5.3 shows the mean and standard deviation of the valence ratings depending on the use of positive or negative typefaces. We discovered that the participants reported higher valence ratings with positive or negative typefaces than with the neutral typeface. To observe each participant's rating with and without typeface, a paired t-test was conducted (at the .05 level). For all the tasks, participants had significantly stronger valence effect with the positive or negative typeface than without it (Negative 1: $p < .001$, Negative 2: $p < .001$, Positive 1: $p = .021$, Positive 2: $p = .044$).

In this section, we conducted a survey study and observed significant differences depending on whether or not the positive or negative typeface was used. Although the contribution of font varies depending on the conversations, we confirmed that the use of fonts changed the meaning of the message. From the results, we can reject the two null hypotheses H_{01} and H_{02} and conclude that the use of typeface in a message can affect and intensify the valence received by users.

Table 5.3 The results of Task 2. The values in each cell indicate MEAN (SD). The participants reported higher valence ratings with positive or negative typeface than with the neutral typeface. The values with * mark indicate a statistically significant difference at the .05 level.

	Neutral	Positive	Negative	p-value
Negative 1*	3.53 (.90)	-	4.42 (.79)	<.001
Negative 2*	3.61 (.76)	-	4.36 (.79)	<.001
Positive 1*	2.83 (0.83)	3.06 (1.03)	-	.021
Positive 2*	2.92 (1.01)	3.31 (1.02)	-	.044

(a)	A: I won the lotto.		
(b)	B: How much did you win?		
(c)	A: 5 dollars!		Neutral
		5 dollars!	Positive
		5 DOLLARS!	Negative
(d)	B: _____		

Fig. 5.7 Example conversation task (3. *Winning the Lotto*)

5.5 Qualitative Study: Exploring User Experiences

For qualitative analysis, we used a focus group study. The study was separated into two parts. First, we designed a role-playing study that pairs of participants interacted with each other in prescribed situations. By doing that, we can not only make all the participants get familiar with the system but also see the emotional effect of the typeface in the semi-structured situation. After that, we equipped a focus group discussion session where two participants who had interacted with each other and an instructor participated as a moderator. By illustrating the actual user experience, we aimed to find the potential value of the proposed system and leverage the insights from the role-playing study by presenting actual user comments in group discussions.

5.5.1 Method

Participants

According to [56], as few as three to six focus groups are likely to identify 90% of the themes to be observed, we recruited five groups of participants. We recruited five pairs (ten participants; six females, four male) of participants who were friends, and made each pair into one group. They were between the ages of 24 and 34 and were of various nationalities, including American, Chinese, German, Indian, Korean, and Taiwanese. They had sufficient experience using a mobile messenger. English was used

Table 5.4 Vignettes for the role-playing test. The line (c) has variations in the typeface used.

1) Magazine Subscription
(a) A: I like this magazine.
(b) B: So do I. It gives you all the news.
(c) A: Listen, I gave a subscription to my parents.
(d) B: _____
2) Olympic Season
(a) A: I've been looking forward to this Olympic season so much.
(b) B: Me too! Excitement each day!
(c) A: Anyway, did you watch the soccer game?
(d) B: _____
3) Winning the Lotto
(a) A: I won the lotto.
(b) B: How much did you win?
(c) A: 5 dollars!
(d) B: _____

as the common language for texting. The session length of the user test was approximately two hours. All participants were compensated with a book voucher worth about US\$20. No users had been exposed to the Emotype application before the experiment. To familiarize users with the system, we provided a short tutorial and then gave them time to use the prototype freely.

Environment

The application was run on an Android 4.4 device (7.02-inch screen, 1920 × 1200 pixels at 323 ppi). We conducted the user test in various contexts, (e.g., a lecture room, coffee house, and public lounge). For the role-playing study, a pair of users sat in separate rooms where they could not see or hear each other. For the focus group study, two participants who had interacted with each other formed a group for the study, and an instructor participated as a moderator.

5.5.2 Role-playing study

Task

We designed three vignettes for the role-playing experiment referring to [1] containing the various dialogue examples (see Table 5.4). The participants in a pair were randomly assigned a role *A* or *B*. Figure 5.7 shows an example conversation task. All three vignettes were performed three times, and at each attempt, *A* was requested to have different feelings and changed the typeface for message (c) (neutral, positive, and negative). Then *B* replied (d). In the experiment, *B* surmised about how *A* felt

Used typeface A	Reported feelings B		
	neutral	positive	negative
neutral	0.47	0.53	0.00
positive	0.13	0.87	0.00
NEGATIVE	0.00	0.06	0.94

Fig. 5.8 How well the recipients B guessed the sender A's emotion depending on the typeface used. The sum of each row is one ($p < .001$ by Fisher's exact test). As we can see the highlighted results (red-dotted boxes), users reported positive feelings with both the neutral typeface and positive typeface with high probability.

on message (c), and was requested to submit the option that gave the closest match between negative, positive, or neutral.

Results

How the Emotion that the Recipient Inferred Changed: We analyzed B's report of the role-playing test. Because five pairs conducted three situations while choosing among three typefaces ($5 \times 3 \times 3$), we investigated 45 user reports and responses.

Figure 5.8 shows how well the recipients B guessed the sender A's emotion depending on the typeface used. We observed that, if a sender used the neutral typeface, the recipients guessed the sender's feeling as being neutral or positive almost half the time. However, if a sender changed the typeface, the reported feelings varied significantly depending on the typeface used. Fisher's exact test shows that the result is significant ($p < .001$).

How the User's Reaction Changed: We analyzed responses that reported the same emotion regardless of the typeface used. From Figure 5.8, it is evident that users reported positive feelings with both the neutral typeface and positive typeface (red-dotted boxes). As far as the user reporting is concerned, it seems that the influence of the positive typeface was not much different from that of the neutral typeface. To investigate the effect of the positive typeface, we examined how B's reaction changed in accordance with typeface usage.

Row (d) in Figure 5.9 shows actual response examples by participants who guessed that the sender's intent was positive regardless of which typeface was used (positive or neutral). By reviewing the actual responses, we observed that the positive typeface elicited the more positive response than that of neutral typeface. We can see that users who take role B tended to use emphasizing expressions much more ("so" (P2), "very" (P4), "such a" (P6)) as responses to a message with the positive

typeface than to those using the neutral typeface. Even more, the positive words appeared more frequently in the emotional typeface task (“*useful*” (P4), “*good*”, “*happy*” (P6), “*cool*” (P8)). Here, we can associate this to the result from $H0_2$ in our preliminary study. The typeface intensified the valence which a message conveys, then the recipients who perceived the intensified emotion tried to express their emotion more actively as responses to the perceived emotion.

5.5.3 Focus group discussion

In the focus group discussion, participants discussed their diverse experiences with regard to given special message examples. In addition to that, we explore values of Emotype via conversations obtained from free chats between participants.

Emotion Words, Emoji, and Typefaces in Texting

We provided message examples to participants which were equivalent in contents, but different in the way expressing emotion. Figure 5.10 shows an example. Then they reported the differences between emotion words, emoji, and typefaces.

Findings: All participants reported that the positive emotion conveyed by neutral fonts (Figure 5.10(3)) evoked formal and business-like feelings. On the other hand, they felt that the speaker genuinely seemed happy in the messages that used emoji and typefaces. “*Both the emoji and typeface conveyed very positive feelings, and the message with the typeface sounded more intimate*” (P4). The comparison between typeface, emoji, and emotion word usages revealed that both typeface and emoji conveyed emotion effectively, but the nuance between them was slightly different. “*With the emoji, I felt the sender was smiling after she or he says the words, but with the typeface, I felt the sender was saying it in a happy tone of voice all along*” (P6).

We surmise the observations comes from inner voice experience [62] — speech rehearsal in one’s mind. In other words, when people read the message with typeface, the unique visual appearance affects their inner voice experience and then would give them feelings such as a *happy tone of voice*. On the other hand, in case of negative messages, nonverbal signals in emoji and typeface eased the negative mood in the chat. All participants mentioned that negative feelings conveyed by text only evoked more negative feelings than the emoji and typeface. “*If I were really in negative mood, I would not use emoji or typefaces. But if I didn’t want to make others worry about me, a message with the typeface would seem to be effective*” (P5). Some participants also reported that they felt a tone of voice that guides them to speaker’s personality with Emotype: “*The message with negative typeface sounded like someone’s voice who has a violent temper*”(P6).

Inconsistency between the Content and Typeface

It is known that emotional conflict tasks draw unusual behaviors because of the implicit emotion regulation [58]. We therefore investigated how inconsistencies between the content and the typeface

(a)	A:	I like this magazine		→ (P1/3/5/7)
(b)	B:	So do I. It gives you all the news.		→ (P2/4/6/8)
(c)	A:	“Listen, I gave a subscription of a magazine to my parents.”	“ listen. i gave a subscription of a magazine to my parents. ”	→ (P1/3/5/7)
(d)	B:	Nice, how do they like it?	So nice ;) they must like it	→ (P2)
		Oh, that's great. Did they like it?	Nice! ;) did they like it? I guess it is very useful for them	→ (P4)
		Subscription?	You are such a good girl! They must be happy about it!	→ (P6)
		Why?	That's cool	→ (P8)

Fig. 5.9 The actual user responses in the vignettes 1. *Magazine Subscription*. We observe that the use of the positive typeface elicited an active response in row (d). We highlighted emphasizing expressions in bold, and colored the positive words green. For example, in row P1, with the neutral typeface, P1 simply gave a positive response (*Nice*), and asked (*how do they like it?*). However, with the positive typeface, P1 showed a more active positive response (*So nice ;)* and affirming (*they must like it*). Here, P_i indicates the id of each participant ($i=1,2,3,\dots, 8$).

- (1) Wow, that's great! 😊
- (2) **wow. that's great!**
- (3) Wow, that's great! I'm glad to hear that.

Fig. 5.10 An example using three different types of emotional expression: Emoji, Emotype and emotion word.

	Content	Typeface		Inconsistency
(1)	Positive	Positive	i am happy these days!	
(2)	Positive	Negative	I AM HAPPY THESE DAYS!	✓
(3)	Negative	Negative	I GOT FIRED	
(4)	Negative	Positive	i got fired	✓

Fig. 5.11 Messages showing inconsistency between the content and the typeface.

affected the participants' feelings and thoughts. All participants received four short messages (Figure 5.11) and were asked to report how they felt in response to the messages. The content and typeface of message (1) were both positive; however, message (2) had positive content and negative typeface, therefore showing inconsistency. In the same way, the message (4) also shows inconsistency.

Findings: One of the functions of messages that showed inconsistency between the content and typeface was humor in a sarcastic situation. According to participants P1 and P2: *"It reminded of me the TV animation, South Park, in that expression of a sense of humor."* In addition to P1 and P2, all other participants mentioned that they experienced a humorous feeling from the sarcastic situation. It seemed that the inconsistency in a single media only caused confusion; however, the inconsistency between the two media, e.g., a positive text but negative typeface, was perceived as a humorous mood by the recipient. Some participants described it as follows: *"If I received the message [I am happy and angry], it only makes me confused. But with the message with (2) [I am happy (with anger font)], it seems funny. It seems like the sender is trying to make the mood humorous even though he or she was really upset"* (P5, P6). The use of typeface also enabled participants to imagine the situation in detail: *"From the message, [I got fired] with a happiness typeface, I imagined a situation in which the sender was fired from a job she or he really disliked"* (P7, P8).

These reports implied that the inconsistency between the text and typeface was perceived by users as humorous feeling and the use of typeface contributes to rich emotional experiences.

Dynamic Tone Change in Free Chat

It is known that seriatim transmissions of parts of a message are characteristic of instant messages (e.g., hey man [send] what's up [send]) [15]. In the free chats between participants, we discovered that this characteristic brought about interesting Emotype usage.

Findings: Figure 5.12 shows an interesting example of usage. Participants preferred to express different emotional signals in accordance with the segment of the message. They actively utilized

P2(1): I want to quit studies and open an Indian restaurant
 P2(2): how do u think about the idea?
 P2(3): it's my new goal
 P1(1): ~~SORRY, BUT I THINK THIS IS A STUPID IDEA~~
 P1(2): one day it will pay off

Fig. 5.12 An free chat example which shows dynamic tone changes.

typeface changes in seriatim messages. This usage fulfilled the need to disclose the sender's feeling or intention. P1 explained why he showed a dynamic tone change between the messages P1(1) and P1(2), "*I used the positive typeface to warn about the negative future that P2 may have, but not to make the mood too serious*" (P1). This kind of usage also observed in other participants' conversations.

As we observed in the free chats, the dynamic usage of typefaces in seriatim transmissions of messages enabled the expression of multiple emotions in a message and conferred a lively mood in the conversation. Furthermore, typefaces were being used naturally to create a relaxed atmosphere to the chat.

5.6 Discussion and Conclusions

This study has proposed an approach that employs typefaces to convey intended emotional states in a mobile environment. To examine the effectiveness of this approach, we designed user studies. In our preliminary study, we demonstrated the feasibility of using typefaces to communicate emotions. If the target message was ambivalent, the perceived emotion varied greatly depending on the typefaces used. This result indicates that emotions are being transmitted through another channel—the typeface, even if the emotions are not clearly stated in the text. If the emotion of a message was explicitly mentioned, the use of a typeface that matched the emotion of the message emphasized the emotion. In the user study, we explored the user experiences with role-playing and focus group discussion studies. We observed that the use of typeface not only modulates the emotional signals a message conveyed, but also elicits active responses in expressing emotions. We also obtained various user experience reports. We observed that there was a slight difference in nuance between typeface and emoticon, and participants reported this as acoustic experiences. The inconsistency between the content and typeface created a humorous feeling and enabled participants to imagine the situation in detail. In the free chats between participants, we observed that they naturally exploited typeface to guide a conversation toward an intended mood. Here, typeface contributed to create rich experiences such as dynamic tone changes. The series of comments from the focus group discussions suggested that the Emotype conveys nonverbal signals, such as a tone of voice. In other words, Emotype was demonstrating “how

it was said” in the chats. Our findings indicate that the use of typefaces in mobile communication will expand channels of nonverbal signals and contribute to users’ mobile communication.

However, several limitations remain. We mainly investigated the emotional effect of two typefaces, negative and positive. The emotional effects of other typefaces need to be examined. Another issue comes from individual differences. Even if there was a strong consensus on the emotional effect of a given typeface, these effects could vary according to the user. Finally, we could not observe user behaviors in natural situations. Because the participants knew that conversations will be analyzed, it was difficult for them to chat naturally. This eventually led our user study to be conducted in the limited situation. We hope that Emotype will be applied to real messenger applications, and expect to find further values through a long-term study like emoji/emoticon researches have done.

Chapter 6

Font Search by Image based on Color-based and Concept-based Matching Methods

One of the important aspects in graphic design is choosing the font of the caption that matches aesthetically the associated image. To obtain a good match, users would exhaustively examine a long font list requiring them a substantial effort. To address this issue, we present two font search systems that enable users to use images as queries - (1) query by image impressions based on color study and (2) query by image contents based on concept analysis. Instead of matching font and image directly, we perform mapping both image and font to color-based semantic space or concept-based semantic space. Our evaluation results show that the recommended fonts scored better than other comparisons and provides competing results with the ones chosen by experienced graphic designers.

6.1 Introduction

In digital graphic design projects, designers use various elements such as background images and typography elements. Unity in design makes these elements come together as a whole in order to communicate a particular message. In other words, to create effective graphic works, users should consider impressions that each element conveys and should make all the elements match well with each other. Figure 6.1 shows an example of matching two fonts that give different impressions with an image. The image may convey soft and delicate feelings from the presence of flowers and soft color tones, and the font on the left matches better this feeling as it has a light weight that appears to be carefully written. However, finding a best matching font to a given image is a difficult task.

The most general way to search font is to perform an exhaustive font search through a list of fonts that may match well the given image. This is a time-consuming and laborious process for novice users as well as professional designers. According to [134], providing too many choices will not only stress the users but will also decrease their satisfaction and enthusiasm in performing the font selection. Addressing this issue, this chapter presents two font search systems which lead to an efficient and satisfactory decision-making for the font selection.

This chapter proposes two systems that enable users to use an image as an input for the people who have difficulty in finding fonts that match the image input well. However, there is neither image-font pair dataset nor direct guidance on font image matching. Instead of matching font and image directly, this study rather performs mapping both image and font to the same semantic space. Proposed two systems use the following mapping spaces which are designed for dissimilar purposes (Figure 6.2b).

- **Affective color-based semantic space:** Color features are one of the most effective media for expressing feelings and emotions [152]. We map both font and images to the color-based semantic space consisting warm-cool and hard-soft axes (Figure 6.2a).
- **Affective concept-based semantic space:** What is depicted on an image affects the perceived impression of the image. The proposed model firstly predicts what concepts are depicted in the image. Then the concepts are mapped to vectors of real numbers using word vector space models. In the same way, tags of a font are mapped to vectors of real numbers (Figure 6.2b).



Fig. 6.1 The flower image with soft color tones conveys soft and delicate feelings. We can say that the thin serif font on the left matches well the image as it gives more similar feelings than the thick and angular font on the right.

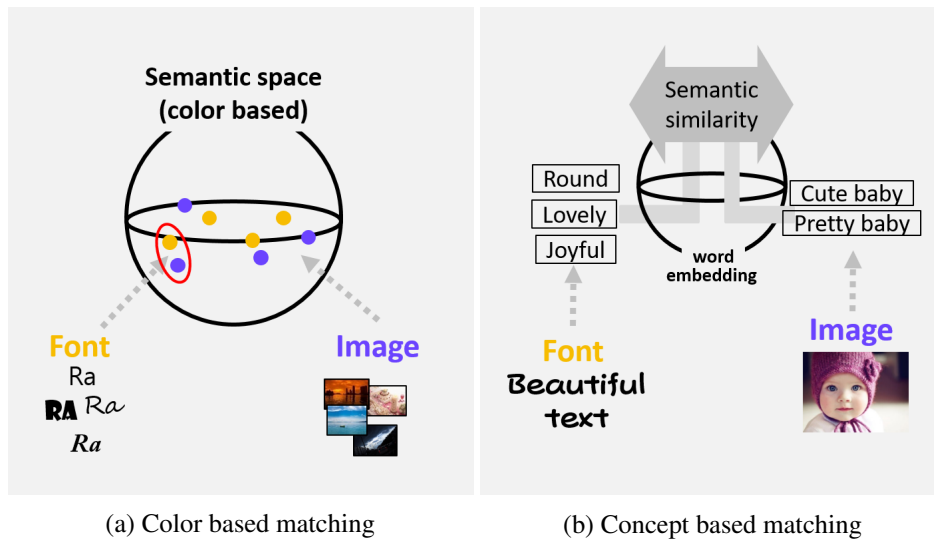


Fig. 6.2 Proposed two font search methods that use two different semantic spaces.

The remainder of this chapter is organized as follows. Section 6.2 reviews the related work on font search. The methodologies for color-based and concept-based font image matching is then illustrated on Section 6.3 and 6.4 respectively. In Section 6.5, we discuss about the limitations of each system, and then conclude in Section 6.6.

6.2 Related Works

6.2.1 Font search

There are various approaches for searching fonts. Font sharing web sites enable users to browse fonts by categories such as serif and script, and properties (e.g. thickness of letters) [38, Adobe]. A recent visual font recognition study enabled the identification of fonts from real-world text images such as signboard [154]. These systems are only useful if a user has a clear preference for the fonts to be searched. IDEO proposed a tool that uses machine learning to learn the visual similarity of fonts and allows their visualization in a two-dimensional latent space [IDEO]. O’Donovan et al. proposed an interface for exploring fonts using binary attributes such as “dramatic” or “not dramatic” [111]. Although both studies proposed an interface for exploring large font datasets without a clear preference, their interface still requires a high aesthetic intuition of the user. In this chapter, we propose alternative approaches that enables users to explore fonts that match in an harmonious way the feelings the image conveys.

6.2.2 Word embedding in search

Recently, word embeddings such as Word2Vec [105] attracted attention in IR community. Initial attempts to use word embeddings in IR system is to use it for improving user query, e.g., query

expansion. Word embedding is used to find a semantically related terms to the original query for query expansion. Using word embedding for query expansion improved performance by solving the tag-mismatch problem [90, 43]. Kusner et al. proposed word embeddings based distance metric to calculate dissimilarity between two text documents [89]. Even though word embedding methods have achieved a great success in many natural language processing tasks including query expansion, document similarity and so on, there is a large gap between the learning objectives of word embedding [105] and general IR tasks [170]. Motivated by the challenge, Zamani et al. proposed relevance-based word embedding that better understands the purpose of the IR tasks. However, there is still a limit to apply the above methods to real font web community setting in practice. The retrieval result are dependent to parameters such as how many terms to use for tag complement or query expansion. To learn relevance based embedding [170], a very large query-document pair dataset is needed.

6.3 Affective Color-based Matching

We highlight our key contributions as follows. Firstly, we model a font-impression relationship on a two-dimensional semantic space and built a font recommendation system to save users from the effort to explore large font datasets. If a *warm and soft* feeling image was given as an input, the system recommends *warm and soft* feeling fonts. Secondly, our system provides an explanation to support the recommendations made. It has been shown in [64] that providing a recommendation that lacks an explanation, brings about users' lower satisfaction in the recommended results. Explanations why the system recommended the items give the users the feedback that their requests were understood by the system. The explanation our system gives describes what is the important feature of recommended fonts, (e.g., *the cursive shape is the distinguishing feature of the recommended font.*). Finally, users can also provide their feedback by controlling the typography attributes e.g., *more slanted*. The goal of our work is not only to recommend the best matching fonts but also to make users feel satisfied with the search by increasing the design-domain knowledge via an explainable recommendation system.

6.3.1 Methodologies

In this section, we describe our modeling method for mapping fonts on a two-dimensional semantic space, and then introduce two interface features, i.e., how we generate the explanation for supporting the recommendation and how the system processes the user feedback.

Overview

Figure 6.3 outlines the features of FontMatcher and the actual user interface. If a user inputs an image, a content analysis tool [Google] identifies objects in the image (e.g., cake and table) and an impression model [32] analyzes the feelings the image conveys (e.g., soft and tender). Then, a natural language generating engine [51] outputs a sentence describing the image in panel (e.g., *The image with cake and table gives soft and tender feelings*).

The image impression model [32] outputs a coordinate in the semantic space of the input image. Because our font impression model analyzes fonts in the same semantic space, coordinate values are used for calculating semantic distances between the image and each font in the semantic space. Then the system ranks fonts in ascending order by their distance and displays the result in panel 2. At the same time, the system extracts distinguishing features of top-ranked fonts and generates a description in panel 3, e.g., *Small x-height, very light weight and very cursive shape are distinguishing features of the recommended fonts.*

Users can give feedback in panel 4 by controlling six font appearance features, such as slope, x-height, alignment, shape, weight contrast, and weight. For example, if a user wants to find more slanted fonts than the current recommendation, he/she can select the *more slanted* button in panel 4 and the results are updated in panel 5.

Font Impression Mapping

1) Mapping Space Kobayashi developed a semantic space consisting of warm-cool and hard-soft axes [84]. He presented two semantic spaces, word and color. On each space, 180 impression words or 1,170 color combinations were located respectively. The bottom right of Figure 6.3 shows four semantic spaces: font, word, color and image. According to Kobayashi, different objects located in a certain position in different spaces have a similar feeling. He mentioned that not only color, but other objects such as clothes and house designs can be mapped in the semantic space as well. We, thereby, utilize a model which maps natural photographs on the image semantic space [32], and furthermore, we built a model for font mapping on the semantic space to correlate images with fonts.

2) Font dataset O'Donovan et al. provided a 200 fonts dataset and 37 attribute scores for each font [111]. We dropped three fonts which have inadequate shapes (e.g. too wide to display in a limited space). We then select a subset out of the provided 37 attributes for model construction. We removed 7 typography limited attributes (e.g. capital, display, italic, etc.) and 18 attributes which were not defined on the semantic space (e.g. angular, attention grabbing, and so on). Consequently, we used 197 fonts with 12 attribute scores in our mapping process.

3) Mapping method To map fonts with 12-dimensional attribute scores into a two-dimensional semantic space, we adopt a dimensional reduction method. There are many dimensional reduction methods such as principal component analysis (PCA) and multidimensional scaling (MDS). However, the reduced axes by these methods are not explainable. We, therefore, applied factor analysis that explains the extracted factor with related variables for the modeling process.

The factor analysis of the 12 attributes yielded three factors which explain 85.19% of the variance. Table 6.1 shows the factor analysis results with varimax rotation. We can find that the first factor comprised of five attributes describes *coolness*. The second factor comprised of four attributes describes *warmness*. Three attributes yield the third factor, and we labeled this as *strong to soft*. We

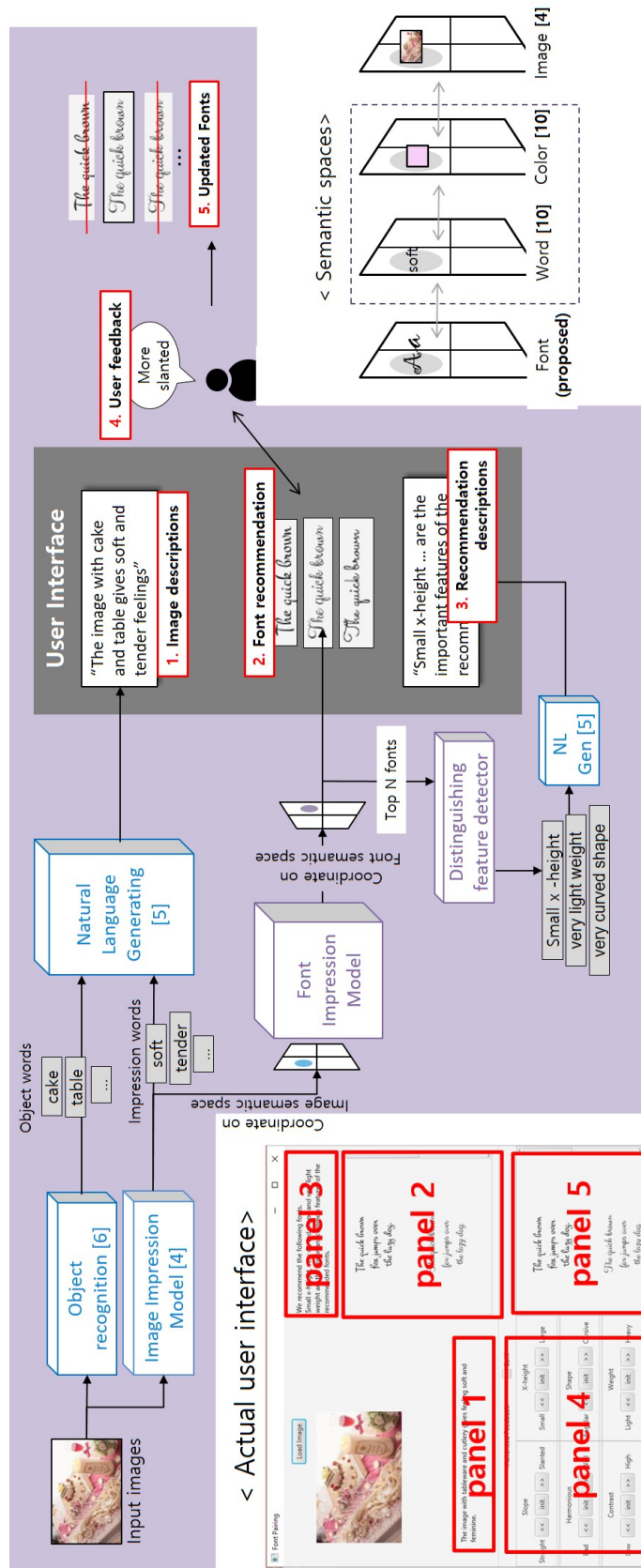


Fig. 6.3 Overview of the system flow. If a user inputs an image, a image description is generated by object recognition [Google], impression analysis [32] and natural language generating models [51] in panel 1. Then our proposed font impression model recommends fonts which have similar feelings to the given image (panel 2). Based on the recommendation, the system generates a description for supporting the recommendation (panel 3). Users can give feedback by controlling font features such as *slanted* (panel 4), and the results are updated in panel 5.

Table 6.1 Factor analysis result. The column α shows Cronbach's alpha.

Attribute	Factor 1 (coolness)	Factor 2 (warmness)	Factor 3 (strong-soft)	α
formal	0.892	-0.015	-0.029	.896
gentle	0.803	0.339	0.354	
calm	0.762	-0.032	0.579	
fresh	0.744	0.449	-0.145	
sharp	0.719	0.21	0.327	
friendly	0.599	0.654	0.185	.929
charming	0.091	0.948	0.069	
happy	0.032	0.919	0.065	
graceful	0.334	0.86	0.004	
delicate	0.116	0.57	0.733	.885
soft	0.442	0.106	0.845	
strong	0.031	0.082	-0.971	

KMO measure of sampling adequacy : 0.880
Bartlett test p-value < .000

used the coefficient matrix obtained from the factor analysis to calculate the 197×3 factor score vector F ($F = X \times B$, where X is the 197×12 attribute score matrix and B is the 12×3 factor score coefficient matrix). According to the semantic space, we can say the third factor F_3 describes the hard-soft axis well. We normalized the 197×1 column vector F_3 , and regarded an element of the vector as a font's coordinate value on the hard-soft axis. In the case of warm-cool axis, we observed two factors (F_1 and F_2) explain the axis (coolness and warmness). We therefore performed a linear combination of the first (F_1) and second (F_2) factor score vectors. By doing that, we can represent the two factors in a single axis Y (warm-cool axis). We formulate the linear combination as: $Y(\alpha) = \alpha F_1 - (1 - \alpha) F_2$. Then we estimated the optimal α as follows.

$$\alpha = \operatorname{argmax}_{\alpha} \sum_{i=1}^5 \frac{w_{cool}^i \operatorname{corr}(Y(\alpha), F_{cool}^i)}{5} + \sum_{j=1}^4 \frac{w_{warm}^j \operatorname{corr}(Y(\alpha), F_{warm}^j)}{4}$$

The first term in the right side indicates the sum of correlation between $Y(\alpha)$ and five attributes' factor scores F_{cool} which describe *coolness*. The second term is for the four attributes' factor scores F_{warm} which describe *warmness*. Here, w_{cool}^i (or w_{warm}^j) is the weight of the attribute i (or j). Each weight value is the value on the warm-cool axis provided by the word semantic space [84]. As a result, we were able to get coordinate values of 197 fonts on the warm-cool axis estimated by the linear combination of the first and second factor scores. Figure 6.4 shows the result of mapping 197 fonts to the semantic space and some example fonts. For example, if a user inputs an image, the image

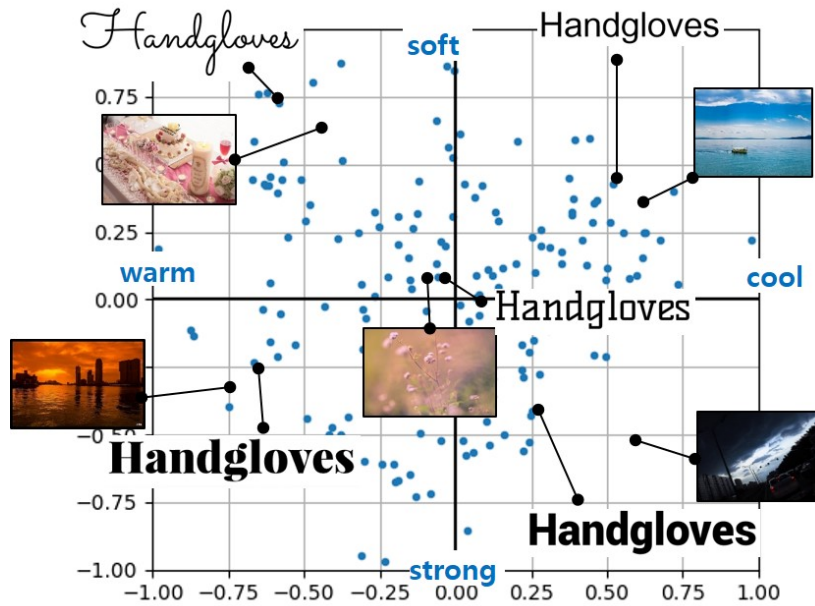


Fig. 6.4 The font mapping results on the semantic space. One blue dot indicates one font. The five images were plotted in the semantic space by the image impression model. The phrase “Handgloves” is a filler text to demonstrate the shape of fonts.

impression model outputs a coordinate value on the semantic space. Based on the value, the system calculates the distance from the image to each font. Then the system ranks fonts in ascending order by the distance and displays the result.

Font Description Generating

Font description comprises of combination of the **font feature word** and *the degree of the feature word*, e.g., *very heavy weight*. We defined six font features which are picked up from [63], i.e., *slope, x-height, alignment, shape, weight contrast and weight*. We also defined five stages of degree for each feature. For example, *weight* has five stages of the degree words, i.e., *very heavy, heavy, normal, light and very light*.

For recommended top N fonts, we extracted six font features and computed each histogram. The same process was conducted for whole font dataset (197 fonts). After that, we measured the distribution difference between the top N fonts histogram and the 197 fonts histogram by using the KL divergence [88]. Because it returns how much the distribution of the recommended fonts is different from the whole font dataset, we regarded the feature which has the largest difference as a distinguishing feature. Here we adopt three features which returned the highest estimated differences for generating descriptions. Referring to the mean feature value of the top N fonts, we assigned *the degree of the feature* word for the font feature. For example, if the mean weight value of top N fonts is 0.9 (five steps of the degree in 0.2 intervals between 0 to 1), we assign the degree word *very heavy*. We

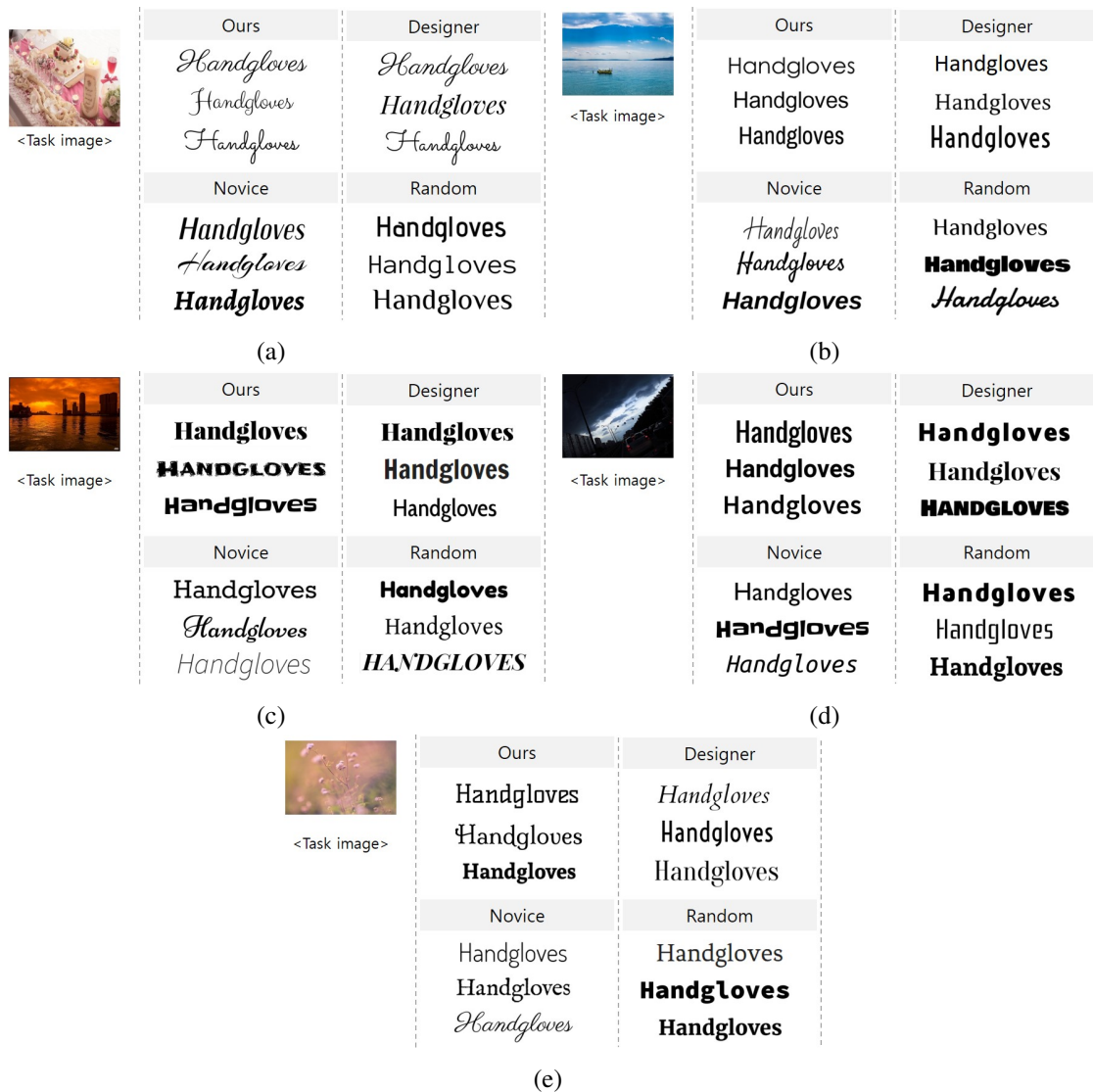


Fig. 6.5 Examples of font-image pairing task results

then generated a description which is a combination of the font feature words and the degree of each feature word (e.g., *Small x-height, very cursive shape, and very light weight are the distinguishing features of the recommended fonts.*

Attribute Filtering

For the given recommended fonts, users can give relative feedback such as “*heavier weight than the recommended one.*” If a user gives feedback *more* than the current weight score, the system calculates the mean of the weight value of the top N fonts as the current feature score, and then filters out the fonts which do not satisfy the constraint from the ranked font list.

6.3.2 Experiment

To evaluate how well the system recommends fonts, we conducted a user study. Four different user groups chose the font that matches in an harmonious way the feelings a given image conveys, then designers examined and scored the image-font pair. In addition to that, the effect of accompanying explanations was also evaluated.

Effectiveness of the System

Figure 6.4 shows five different images we used for the user study. Participants were requested to choose best three fonts which harmoniously match a given image. We formed four user groups, designers, novices, random and system and each group conducted five font-pairing tasks. We recruited three designers and three novice users, so 45 font-image pairs were generated per group (3 participants \times 3 fonts \times 5 tasks). For the random group, instead of involving the human intervention, we randomly chose three fonts per each task arriving at 45 pairs. Image-font pairs of the system group are the result of three top-ranked fonts. In total, we obtained 150 font-image pairs for the test. Figure 6.5 shows font-image pair examples of four different group to the given image.

Effect of Explanation

We organized three user groups via Yahoo crowd-sourcing service to investigate the effect of the explanation on users' reliability of the recommended results. We provided image-font pairs only for the first group (default), image-font pairs with image description for the second group (image description), and image-font pairs with both image description and font description (image & font description). We used the five sets of three image-font pairs provided by our system in the font-pairing task. Workers rated two questions using a five-point Likert scale, (1) how reliable and (2) how satisfying the image-font pair is. Each worker rated five image-font pairs, and 100 workers were allocated per group, so we finally collected 1,500 assessments.

6.3.3 Results and discussion

Effectiveness of the system

Six designers who did not participate in the font-image pair generating task rated the collected font-image pairs, which add up to 838 ratings. Each font-image pair was questioned on a ten-point Likert scale ranging from unacceptable (1) to outstanding (10). Table 6.2 shows the mean, standard deviation, maximum, and minimum of the ratings across all pairs by each group. We can see that font-image pairs designed by professional designers received the highest score. Next came System, Novice, and Random, respectively. We found that although the system recommendation results were not as good as the designer's font-pairing results, the system made better suggestions than both the novice user and random groups. We conducted One-way ANOVA and overall interactions is significant ($F(3,146)=6.91, p<.001$), but Post hoc Turkey HSD test was not significant. One of the reasons is

Table 6.2 The mean, standard deviation, maximum, and minimum of the ratings across all pairs by each group. We can see that font-image pairs designed by professional designers received the highest score. Next came System, Novice, and Random, respectively. We examined maximum and minimum scores of each task and found that Ours has the most stable performance.

	Ours	Designer	Novice	Random
Mean	4.95	5.06	4.82	4.19
(SD)	(1.98)	(1.93)	(2.10)	(1.89)
Max	6.50	6.75	7.20	6.20
Min	4.00	2.33	2.17	2

that the number of samples is not large (45 for Designer, Novice, and Random. 15 for Ours). It seems that our subjective task and evaluation have had a great influence on the results. We examined maximum and minimum scores of each task and found that the score variation is different for each task. Especially it seems that novice users have a different performance by task, and it results in the largest score variation (minimum 2.17 and maximum 7.20). Both Designer and Random showed a large score variation as well. In comparison with others, Ours showed less score variation (minimum 4.00 and maximum 6.50). From this, we can say that the proposed system has reliable performance.

Effect of explanation

We analyzed the assessments in terms of user reliability of the recommended fonts. According to [17], satisfying results positively affect user reliability. Because different workers participated to the three different groups, user reliability was biased depending on the individual's satisfaction. To avoid this bias effect, we defined a relative reliability which measures how the reliability of the recommendation results changed in accordance with the satisfaction of the given results. We labeled 1 (or -1) when a worker's reliability score is higher (or lower) than his/her satisfaction score, or 0 when there was no change. The mean and standard deviation of the average relative reliability scores for the default, image description and image & font description groups were -0.11(.48), -0.07(.50) and -0.026(.52) respectively (the higher the better). A one-way ANOVA revealed the existence of a significant effect of the descriptions on the relative reliability ($F(2,1497) = 3.53, p=.03$). We can see that the image & font description group rated the highest relative reliability. Next came image description and default. This result indicates that the image description and font description which supports the recommendations brought about a higher reliability of the image-font pairs.

User Comments

We invited more novice users and collected their comments. A few examples are listed here: *"The recommended fonts are the exactly same to what I was trying to find"* *"The image description was helpful. When exploring the font list, I was able to compare the description with fonts. It made*

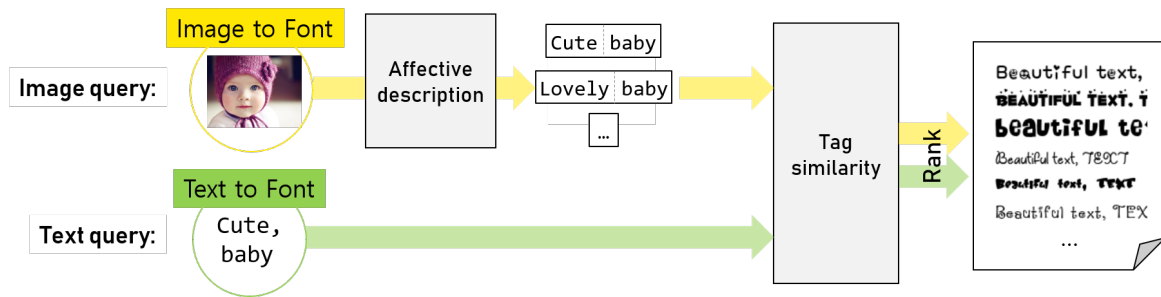


Fig. 6.6 System overview. Users can search font using text query as well as image query.

font-search easy. “I got to know what font features like ‘contrast’ means through the system,” and “This allowed me to think about how to reflect my preference in feedback.” However, some problem was pointed out. “I can check and compare a variety of fonts at once with regular font list, but this system gave the impression that the number of fonts is pretty limited.” “Some of the recommended fonts did not match the description.”

6.4 Affective Concept-based Matching

Font web community is growing. The variety of shapes has increased and the number of fonts is becoming large by social users. As like social image sharing communities, font communities require users to label uploaded fonts with their own keywords, so font search on the web inherits the same problems as social image search has, i.e., tag mismatch [123]. To solve tag mismatch problem, many researchers proposed to use visual content of image [95, 37] or metadata [171]. However, it is difficult to define attribute/class for the font to learn and no metadata exists for the font. To verbalize the visual appearance of a font, people describe the impressions they feel in the font. Even for the same feeling, words they use might vary from person to person. Therefore font tags prone to be more subjective than other multimedia that make the tag-mismatch problem even worse.

We address the problem by leveraging word embedding to relate different words that have similar in meaning. This allows the system to retrieve fonts which are annotated with similar meaning tags, even though those fonts are not annotated with *the user query*.

6.4.1 Methodologies

Overview

Figure 6.6 outlines our framework. Users can use text query as well as image query. If a user inputs text query e.g., cute, baby, our Tag similarity module calculates the similarity between the text query and tags of each font over the dataset. In the case of image query, Affective description module generates affective descriptions e.g., cute baby, lovely baby. Then the estimated descriptions

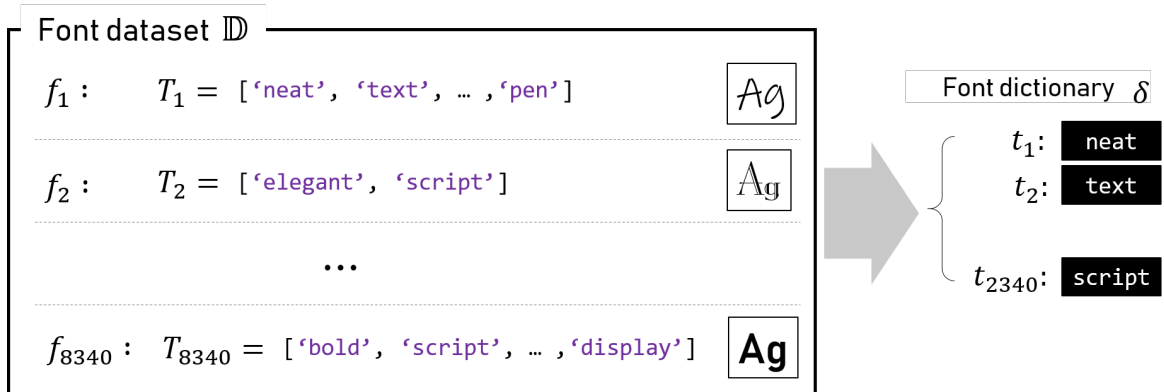


Fig. 6.7 Overview of font dataset. We finally collected total 8340 fonts which has 2,340 size of font tag dictionary.

are fed to Tag similarity module. After that, all the fonts are sorted by the similarity in descending order.

Dataset

Figure 6.7 shows examples of our collected font dataset. We scrapped web pages from font web communities ¹, and collected 9,340 fonts and associated tags. There are 3,712 unique tags across the dataset, and each font was annotated with an average of 8.23 tags. We also excluded words and phrases that are not defined in the word embedding dictionary such as *super hero*. Because dingbat fonts are usually used for describing symbols and shapes instead of letters, we exclude fonts that have the tag *dingbat* and *letterbat* from the dataset. The refined dictionary δ has 2,304 unique tags; $\delta = (t_1, t_2, \dots, t_{2304})$, and the final font dataset \mathbb{D} includes 8,340 fonts; $\mathbb{D} = (f_1, f_2, \dots, f_{8340})$. Each font is averagely annotated with 6.98 tags.

Visual sentiment concepts analysis for image query

The proposed system facilitates multimodal inputs that can use text as well as image query. To understand what concept is depicted in an image, we utilize image recognition model. However, most of the image recognition models are trained with objective information. Because many tags in the dictionary are emotional words, the models are insufficient to our system. Therefore we used DeepSentiBank [26] which is more suitable for our purpose. Based on convolutional neural networks (CNNs), [26] proposed the model to detect 4342 sentiment concepts called adjective-noun pairs that consists of *noun* for visually detectable concepts, and *adjective* for sentiment modulation of the detected *noun* (e.g., *abandoned place*). By using the model, we can get sentiment description a given image, and it is fed to Tag similarity module. Given the query image I , we write top detected

¹<https://www.1001fonts.com/>



(a) Restaurant: long table, fine cuisine, romantic dinner
 (b) Halloween: hill, fake blood, haunted attraction
 (c) Baby: cute baby, happy baby, changing table

Fig. 6.8 Query images and top predicted concepts

M concepts as $Q^I = \{ANP_1, ANP_2, \dots, ANP_M\}$. Here, we separate each adjective-noun pair into two words, e.g., $ANP_1 \rightarrow (q_1, q_2)$ and rewrite the top detected M concepts as $Q^I = \{q_1, q_2, \dots, q_m\}$ where m is twice the M .

Tag similarity

To calculate the similarity between the user query Q^I and each tag set T_k of a font k , we firstly map each word to a vector of real numbers using a word vector space model (it is called word embedding as well). Word embedding is one of the most popular representations of the document. Since it expresses a word as a vector, we can easily measure the meaning different by performing the arithmetic calculation.

With the recent progress in NLP, many word embedding methods are now available. Among the various word embedding, we take advantage of ConceptNet Numberbatch [140] that represents distributional word embeddings such as Word2Vec [105] as well as a knowledge graph that word relationship such as synonyms into account. It showed significantly better performance than others in many evaluations of word relatedness [140]. We used 300-dimensional word embeddings with 417,196 vocabulary size. Given word t , we write the word embedding of the word t as $g(t)$.

The proposed Tag similarity modules calculate the similarity Sim between the user query $Q = \{q_1, q_2, \dots, q_m\}$ and each tag set $T_k = \{t_k^1, t_k^2, \dots, t_k^n\}$ as:

$$\begin{aligned}
 Sim(Q, T_k) &= Sim\left(\left\{t_i^k\right\}_{i=1}^n, \left\{q_j\right\}_{j=1}^m\right) \\
 &= \frac{1}{n \cdot m} \sum_j \sum_i \|g(t_i^k) - g(q_j)\|
 \end{aligned} \tag{6.1}$$

Ranking

Given query Q , the model calculates Sim between the query Q and each tag set of every font in the dataset (T_1, \dots, T_{2340}). After that, all the fonts are sorted by the similarity score in descending order.

Table 6.3 Examples of the search results given image query (Restaurant, Figure 6.8a). We list font names and creators of each font as: 1: Mops by Uwe Borchert, 2: Hot Pizza by Dennis Ludlow (maddhatter_dl@yahoo.com), 3: green piloww by Billy Arget (billyargel@gmail.com), 4: Croissant One Regular by Tipo, 5: Xperience Pasta by Peax Webdesign, 6: Frijole by Font Diner, 7: Dreamwish by Starlight Fonts, Lauren C. Brown (lauren@ork.net), 8: Vanessas Valentine by bythebutterfly, Vanessa (BYTHEBUTTERFLY@GMAIL.COM), 9: Circle Of Love by cutieFont (cutiefont@gmail.com), 10: Snappy Service NF Regular by Nick Curtis. Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.

Rank	Font image	Tag
1	Beautiful text	medium, food, menu, restaurant, diner, serif, headlines, text
2	Beautiful text	famous, pizza, restaurant, food, display, black
3	beautiful text	decorative, fancy
4	Beautiful text	french, food, breakfast, elegant croissant, paris, smooth, serif, restaurant, text, formal, medium
5	Beautiful text	regular, pasta, food, noodles, text, handwritten, thanksgiving
6	BEAUTIFUL TEXT	fried, cooking, spicy, jittery, fat, display, offbeat, food
7	BEAUTIFUL TEXT	cute, romantic, display, stars, dreamy, heavy
8	Beautiful text	light, hearts, romantic, headlines
9	Beautiful text	regular, cute, lovely, playful, romantic, hearts, text
10	Beautiful text	bold, text, american, diner, food, vintage

Table 6.4 Examples of the search results given image query (Halloween, Figure 6.8b). We list font names and creators of each font as: 1: Unquiet Spirits by Sinister Fonts, 2: Hantu Kom Kom by Haslinda Adnan (kakalin2001@yahoo.com), 3: Haunted Eyes Regular by Misti's Fonts (mistifonts.com), 4: Jasper Solid (BRK) by AEnigma (kentpw@norwich.net), 5: JMH CRYPT Regular by Joorge Moron (joorgemoron@gma il.com), 6: Bloodytronic by Brain Eaters Font Co. (info@BrainEaters.com), 7: Raven Song Regular by Sinister Fonts, 8: Metal Macabre by BoltCutter-Design, 9: Casper by DJ-JohnnyRka, 10: Phantom Fingers Regular by Sinister Fonts. Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.

Rank	Font image	Tag
1		regular, ghosts, spirits, ghastly, eerie, narrow, haunted, halloween, pointy, display
2		scary, horror, black, dripping, bloody, ghostly, halloween, display, regular
3		regular, halloween, cursive, connected, horror, scary, grunge, fancy, haunted, script
4		horror, scary, ghostly, ghost, 3d, halloween, display, heavy, outlined
5		horror, bloody, dripping, halloween, black
6		horror, scary, bloody, bleeding, halloween, dripping, narrow
7		regular, tall, display, halloween, creepy, eerie, sinister, spooky, eroded
8		scary, horror, halloween, satanic, display, poster
9		movie, television, horror, halloween, ghosts, fat, poster
10		skewed, halloween, sinister, eroded, ghostly, headlines, dracula, scary, narrow, spooky', 'eerie, heavy

Table 6.5 Examples of the search results given image query (Baby, Figure 6.8c). We list font names and creators of each font as: 1: Sweet Smile by cutieFont (cutiefont@gmail.com), 2: Sunshine Kiddy Font by Merethe Liljedahl, Lime (<https://plus.google.com/114621834783881488812>), 3: Fontdinerdotcom Huggable by Font Diner (<http://www.fontdiner.com>), 4: Fontdinerdotcom Luvable by Font Diner (<http://www.fontdiner.com>), 5: Baby Lexi Medium by bythebutterfly, Vanessa (BYTHEBUTTERFLY@GMAIL.COM), 6: Toyland NF Regular by Nick Curtis, 7: Unicorns are Awesome Regular by Misti's Fonts (mistifonts.com), 8: Addis Ababa by Kimberly Geswein (gesweinfamily@gmail.com), 9: KG Only*Hope by Kimberly Geswein (gesweinfamily@gmail.com), 10: The Happy Giraffe by Misti's Fonts (mistifonts.com). Here, all the fonts are free for personal use. We additionally write the contact information of creators of fonts that are not free for commercial use.








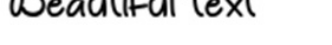
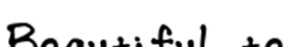

Rank	Font image	Tag
1		sweet, cute, smiley, smiling, kiddy, hearts, text, happy
2		kids, fun, summer, sun, happy, cute
3		fat, poster, whimsical, playful, comic, baby, cuddly
4		fat, poster, cuddly, whimsical, playful, baby, romantic, hearts
5		headlines, handwritten, baby, cute, hairline, thin
6		fat, poster, shadowed, outlined, kiddy, children, cute, toys, toddlers
7		regular, whimsical, handwriting, handwritten, cute, girly, fantasy, unicorn, baby, kids, childish
8		handwritten, whimsical, baby, kids, childish, playful, dotted, cute, handwriting, bold, text, regular
9		cute, quirky, playful, fun, display, regular
10		handwriting, handwritten, cute, happy, adorable, rounded, neat, text, whimsical, medium

Table 6.6 Examples of the search results given image query (Restaurant, Figure 6.8a). We list font names by rank. 1: PopJoyStd-B, 2: RodinWanpakuPro-DB, 3: RodinWanpakuPro-B, 4: RodinWanpakuPro-M, 5: TsukuARdGothicStd-E, 6: TsukuARdGothicStd-B, 7: TsukuARdGothicStd-R, 8: TsukuARdGothicStd-D, 9: TsukuARdGothicStd-L, 10: TsukuARdGothicStd-M.

Rank	Font image	Tag
1	美しい文字	モダン, かっこいい, 大人っぽい, シャープ, 粋, すっきり, 美しい, 繊細
2	美しい文字	モダン, かっこいい, 大人っぽい, シャープ, 粋, すっきり, 美しい, 繊細
3	美しい文字	モダン, かっこいい, 大人っぽい, シャープ, 粋, すっきり, 美しい, 繊細
4	美しい文字	モダン, かっこいい, 大人っぽい, シャープ, 粋, すっきり, 美しい, 繊細
5	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目
6	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目
7	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目
8	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目
9	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目
10	美しい文字	やさしい, 古風, スタンダード, モダン, 色っぽい, シック, 真面目

Table 6.7 Examples of the search results given image query (Halloween, Figure 6.8b). We list font names by rank. 1: ManyoKoinLargeStd-B, 2: ManyoKoinStd-B, 3: ComicMystery Std-DB, 4: MysteryStd-DB, 5: PopFuryStd-B, 6: TsukuBOldMinPr6-R6 SlumpStd-DB, 7: SlumpStd-DB, 8: arcStd-R, 9: RodinHappyPro-EB, 10: RodinHappyPro-UB.

Rank	Font image (font name)	Tag
1	美しい文字	素朴, 渋い, ストイック, おどろおどろしい, 幽霊, こわい, 個性的, 恐怖, 古風, 親しみ
2	美しい文字	素朴, 渋い, ストイック, おどろおどろしい, 幽霊, こわい, 個性的, 恐怖, 古風, 親しみ
3	美しい文字	ポップ, 個性的, 恐怖, 荒々しい, コメディ, 古風, 遊び心, ホラー, おどろおどろしい, 粋, こわい
4	美しい文字	遊び心, ホラー, おどろおどろしい, 粋, こわい, ポップ, 個性的, 恐怖, 荒々しい, コメディ, 古風
5	美しい文字	ポップ, 個性的, 恐怖, 激しい, ファンキー, 荒々しい, コメディ, 古風, こわい
6	美しい文字	可憐, こってり, ゴージャス, オールド, こわい, しなやか, 個性的, 恐怖, 古風, 色っほい, シック, 迫力, 重厚感
7	美しい文字	ゆったり, シンプル, 迫力, 重厚感, やさしい, 萌え, こってり, キュート, やわらかい
8	美しい文字	寂しくなる, フェミニン, やさしい, 可憐, 萌え, やわらかい, 個性的, ピュア, 癒し, あっさり
9	美しい文字	ポップ, 無邪気, 楽しい, 元気, かたい, 重厚感, やさしい, 子どもっほい, 幼い
10	美しい文字	ポップ, 無邪気, 楽しい, 元気, かたい, 重厚感, やさしい, 子どもっほい, 幼い

Table 6.8 Examples of the search results given image query (Baby, Figure 6.8c). We list font names by rank. 1: *UDKakugoLargePr6-EL*, 2: *UDKakugoSmallPr6-EL*, 3: *UDKakugoLargePr6-UL*, 4: *UDKakugoSmallPr6-UL*, 5: *TsukuMinPr5-B*, 6: *TsukuMinPr6-RB*, 7: *TsukuMinPr6-D*, 8: *TsukuMinPr6-LB*, 9: *TsukuMinPr6-L*, 10: *TsukuMinPr6-M*.

Rank	Font image	Tag
1	美しい文字	ゆったり, へたかわいい, ポップ, 軽やか, 無邪気, 楽しい 元気, 子どもっぽい, 幼い, やわらかい
2	美しい文字	かわいい, 素朴, 子どもっぽい, 幼い, 楽しい スタンダード, モダン, あっさり
3	美しい文字	かわいい, 素朴, 子どもっぽい, 幼い, 楽しい スタンダード, モダン, あっさり
4	美しい文字	かわいい, 素朴, 子どもっぽい, 幼い, 楽しい スタンダード, モダン, あっさり
5	美しい文字	フェミニン, やさしい, 可憐, かわいい, 嬉しくなる, 女性的, やわらかい, 癒し, 色っぽい, シック
6	美しい文字	フェミニン, やさしい, 可憐, かわいい, 嬉しくなる, 女性的, やわらかい, 癒し, 色っぽい, シック
7	美しい文字	癒し, 色っぽい, シック, フェミニン, やさしい, 可憐 かわいい, 嬉しくなる, 女性的, やわらかい
8	美しい文字	フェミニン, やさしい, 可憐, かわいい, 嬉しくなる, 女性的, やわらかい, 癒し, 色っぽい, シック
9	美しい文字	フェミニン, やさしい, 可憐, かわいい, 嬉しくなる, 女性的, やわらかい, 癒し, 色っぽい, シック
10	美しい文字	フェミニン, やさしい, 可憐, かわいい, 嬉しくなる, 女性的, やわらかい, 癒し, 色っぽい, シック

6.4.2 Experiment

Example results

Table 6.3-6.5 show examples of the search results given image queries. Each table shows the top 10 retrieved fonts and associated tags. Given the image that looks like a fine restaurant (Table 6.3), Affective description model outputs long table, fine cuisine, romantic dinner (we set M as 3 to use only top 3 affective descriptions). We can find the fonts which include dining-related tags such as restaurant, diner, food or tags for describing the mood such as elegant, formal, romantic ranked high. Even if a font does not have any query words as its tag, the system gave high similarity score to the font that has tags which are semantically related to given query words. Table 6.4 shows the retrieved results of the given zombie image. We can see that the top retrieved fonts have blood dripping or sharp-ended visual characteristics which cause fear to us. The tags of retrieved fonts include ghost, eerie, halloween etc., and those are highly related concepts to the input image. Table 6.5 shows the retrieved results of the given baby image. We can see many fonts have unique visual symbols such as smiley face (font 0), sun (font 1), and hearts (both font 0 and 3). The strokes of those fonts are smooth curves, and the shape of a letter does not fit in the square grid, which gives freewheeling impressions. All the visual characteristics are highly related to the analyzed concepts cute baby, happy baby, changing table, and the tags of the retrieved font reflect this tendency well such as sweet, kiddy and adorable.

In addition to the collected English font dataset, we tried the same task with a small size of Japanese font dataset. The font dataset is provided by Fontworks Inc.². The total number of fonts is 320, and there are 120 unique tags across the dataset. Table 6.6-6.8 show examples of the search results given image queries. In case of the restaurant query image, top-ranked fonts have neat-featured shape. Because Japanese font set has no related tags to the concept dining, the fonts that have the tags for describing the mood such as sharp, beautiful, classic ranked top. In the case of zombie image (Table 6.7), the fonts that have trembling strokes has retrieved. Because the size of font dataset is too small, the fonts after rank 5 do not seem to be related to the concepts of the image. We can find the top-ranked fonts for the given baby image have soft and rounded strokes that create cute feelings (Table 6.7).

Evaluation

To evaluate the proposed system, we invited participants to score the retrieved fonts. We use movie posters for our evaluation. Because movie posters are visual mediums designed to deliver a specific message, background images and fonts in movie posters should come together as a whole in order to communicate a specific message. We evaluate whether the top retrieved fonts are as good as the original font of the movie poster.

²<https://fontworks.co.jp/>

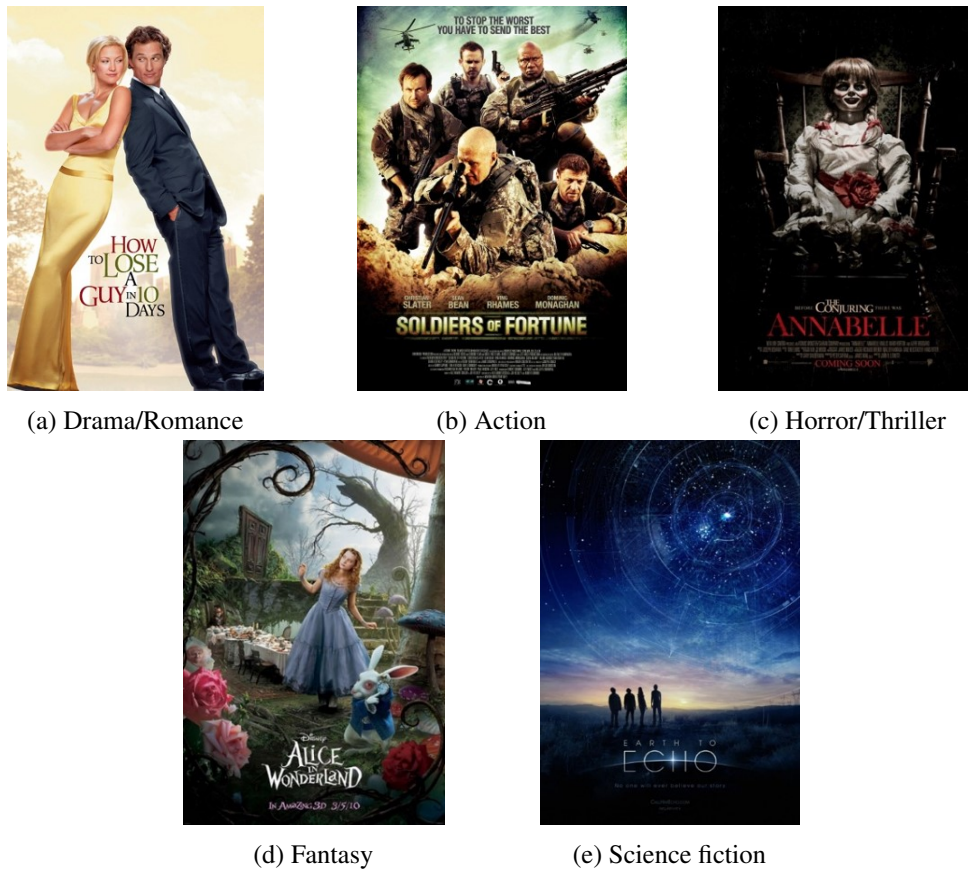


Fig. 6.9 Five movie theme posters (a) *How to Lose a Guy in 10 Days*, (b) *Soldiers of Fortune*, (c) *Annabelle*, (d) *Alice in Wonderland*, (e) *Earth to Echo*

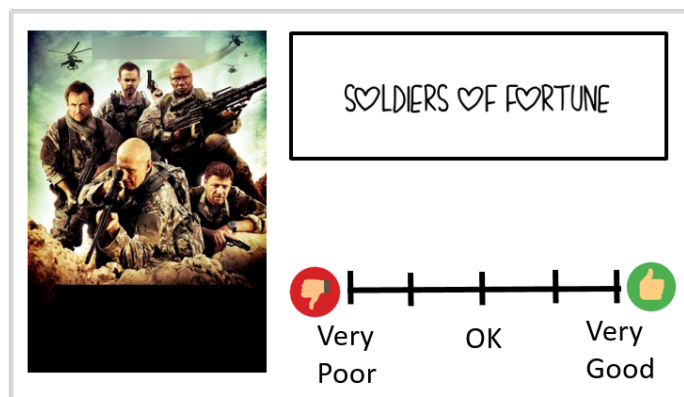


Fig. 6.10 An example task

Table 6.9 Average scores of movie poster tasks. The score ranges from -2 (worst) to 2 (best).

	Original	System	Random
Mean(SD)	0.290 (1.160)	-0.037 (1.270)	-0.470 (1.160)

Table 6.10 Max scores of each movie poster task. The score ranges from -2 (worst) to 2 (best).

		Max	Max scored font image
Romance	Original	0.78	HOW TO LOSE A GUY IN 10 DAYS
	System	0.70	<i>How To Lose A Guy In 10 Days</i>
	Random	-0.16	HOW TO LOSE A G...
Action	Original	0.22	Soldiers Of Fortune
	System	1.05	SOLDIERS OF FORTUNE
	Random	0.19	SOLDIERS OF FORTUNE
Thriller	Original	-0.73	Annabelle
	System	1.00	<i>ANNABELLE</i>
	Random	0.04	Annabelle
Fantasy	Original	0.68	<i>ALICE IN WONDERLAND</i>
	System	1.06	<i>Alice In Wonderland</i>
	Random	0.12	ALICE IN WONDERLAND
Sci-Fi	Original	0.49	EARTH TO ECHO
	System	-0.24	EARTH TO ECHO
	Random	0.21	<i>Earth To Echo</i>

Task: Figure 6.9 shows the five task movie posters. Every movie posters are selected from different themes, Action, Drama/Romance, Science fiction, Horror/Thriller, and Fantasy. Figure 6.10 shows an example of task. We request participants to score how much the font matches well to the given movie poster in five scales. There are total eleven variations of fonts given a single movie poster — one original font of the movie poster, five top-ranked fonts, and five randomly selected fonts. We recruited 49 participants and every participant completed 11×5 tasks.

6.4.3 Results and discussion

Table 6.9 shows the average scores of the collected tasks. The original movie poster font received the highest score. Next came System, and Random respectively. We can find even the average score of system recommendation fonts is lower than that of the original, but it is higher than random. Table 6.10 shows max scores of each movie poster tasks. In the font search, if there was a font that got the highest score among candidates, we can say the search was successful. We, therefore, investigate

which font got the highest score by methods (original, system, and random). In the case of romance movie, one of the top 5 got 0.7 which are very close to the original font score. For action, thriller, and fantasy movie poster tasks, participants gave the highest score to one of the fonts of the system top 5. However, the system search results scored lowest in case of science fiction. Actually, most of the top-ranked fonts were decorated with visual symbols such as stars and clouds, and it seems that the participants thought overly decorated fonts are not suitable for communicating the message of movie posters.

6.5 Limitations

The color-based method has several limitations. The number of fonts user can access is small, and the variety is much lower compared to font sharing web communities. To map a new font to color-based semantic space, all the attribute scores for the font need to be collected. This requires high cost, so this method is not applicable to the dynamic font dataset such as social sharing web communities where the number of fonts increases frequently.

In the case of the concept-based method, it is compatible with any new fonts once the new font is annotated with several words. However, it is sensitive to the accuracy of the affective concept descriptor. If the concept descriptor failed to recognize the input image correctly, it can have a bad influence upon the search results. To improve the search results, the system should be able to cope with user feedback by allowing users to modify the analyzed concepts.

6.6 Conclusions

In this Section, we presented two font recommendation systems for harmonious digital graphic design, color-based, and concept-based search systems. In the color-based system, we built a model for font mapping on a two-dimensional color-based semantic space. Combining the font mapping model and the existing image impression model, the system ranked fonts by a semantic difference between the input image and fonts in the color space. The system also generated explanations to support the recommendation by a statistical method. From the user study, we found that the proposed system made a better suggestion than novice users. Furthermore, the accompanying explanations brought about high reliability on the recommendation. In addition to that, we proposed the concept-based font search system. To understand what affective concept is depicted in an image, we used the CNNs based sentiment concept detector. Then the concepts are mapped to vectors of real numbers using word vector space models. We also mapped the tag set of fonts using word vector space models. By calculating the distance between the word vectors of the sentiment concepts and that of tags of the font, we get the similarity score. From the examples of the search results, we found that the proposed method showed promising results. We conducted a user study to evaluate the system and observed that the system recommended fonts got high scores as good as the professional designer did,

or sometimes better than them. From these results, we believe that the system can be a good starting point for searching fonts in designing graphics.

Chapter 7

Assist Users' Interactions in Font Search with Unexpected but Useful Concepts Generated by Multimodal Learning

When searching for suitable fonts for a digital graphic, users usually start with an ambiguous thought. For example, they would look for fonts that are suitable for a personal web page or party invitations for children. Their design concept becomes clearer as they interact with external interventions such as exposure to suitable images for use in their web page or the children's preferences regarding the party. Hence, it is important to support users' interactions with unexpected but useful concepts during their search. In this section, we present a novel framework that helps users to explore a font dataset using the multimodal method and provides unexpected but useful font images or concept words in response to the user's input. We collect a large font dataset and the associated tags and propose the use of unsupervised generative model that jointly learns the correlation between the visual features of a font and the associated tags for the creative process. By examining the results of the model that change with various inputs, we observed that the model produces highly promising results that appeared to be useful for inspiring users. In the user study, we verified that users actually produce more concepts when inspired by the generated results.

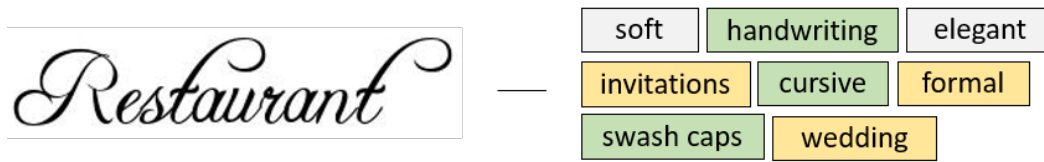


Fig. 7.1 Example of font image and associated tags. It consists of heterogeneous subjective information — the impression the font creates (gray), its application (yellow), and the typographical feature of the font (green).

7.1 Introduction

Inspiration plays an important role in the creative mental process [57]. When we perform creative activities such as painting or writing, inspiration greatly helps us to realize what we imagine. Several artists derive inspiration from external factors such as meeting a variety of people, watching movies, reading books, or paying more attention to their daily life. In this manner, inspiration comes to us when we interact with external interventions.

Owing to technological advances in computer vision and natural language processing, search systems that use natural language [71] or images [54] as an input are available. These systems meet the users' needs for searching for an image using a clear query comprising keywords or images. However, it is difficult to directly apply this conventional search scenario to the creative process as a user generally performs a creative task based on an ambiguous thought that is not sufficiently detailed for a conventional search. For example, in a scenario comprising the searching for a font for a digital graphic, the users' ideas would become clearer in their mind as they interact with external interventions such as exposure to several concepts instead of searching for a specific font.

In this study, we propose a framework that uses machine learning techniques as a tool for the creative process. The results of this study are specifically applicable to users who search for fonts. Typeface designers create fonts to have a specific concept such as impression (e.g., soft or elegant) and its application (e.g., invitations or formal) although each font does not explicitly exhibit these concepts (see Figure 7.1). In other words, tags for font are likely to be subjective. Therefore, even though the reliability of the dataset could be guaranteed, it is likely to be noisy owing to its subjectivity.

In acknowledgment of this problem, the proposed framework learns correlations across various modalities, visual font images, and associated tags in order to become capable of handling a noisy font dataset. For example, given the tag *elegant*, the model outputs the tags *fashion*, *graceful*, *invitation*, or *cursive* that appear to be related to the input tag. We regard these outputs as unexpected but useful, as they reflect the concept of inspiration, and encourage users to refer to the outputs as part of their creative process.

Several studies have been conducted on machines and creativity. Synthesizing a new image that imitates the style of a masterpiece painting from a source image [65, 52] and generating meaningful images that have not been observed in the dataset [73, 124] are representative works of the computer vision community. These works are based on the question, “Can machines be creative?” As compared

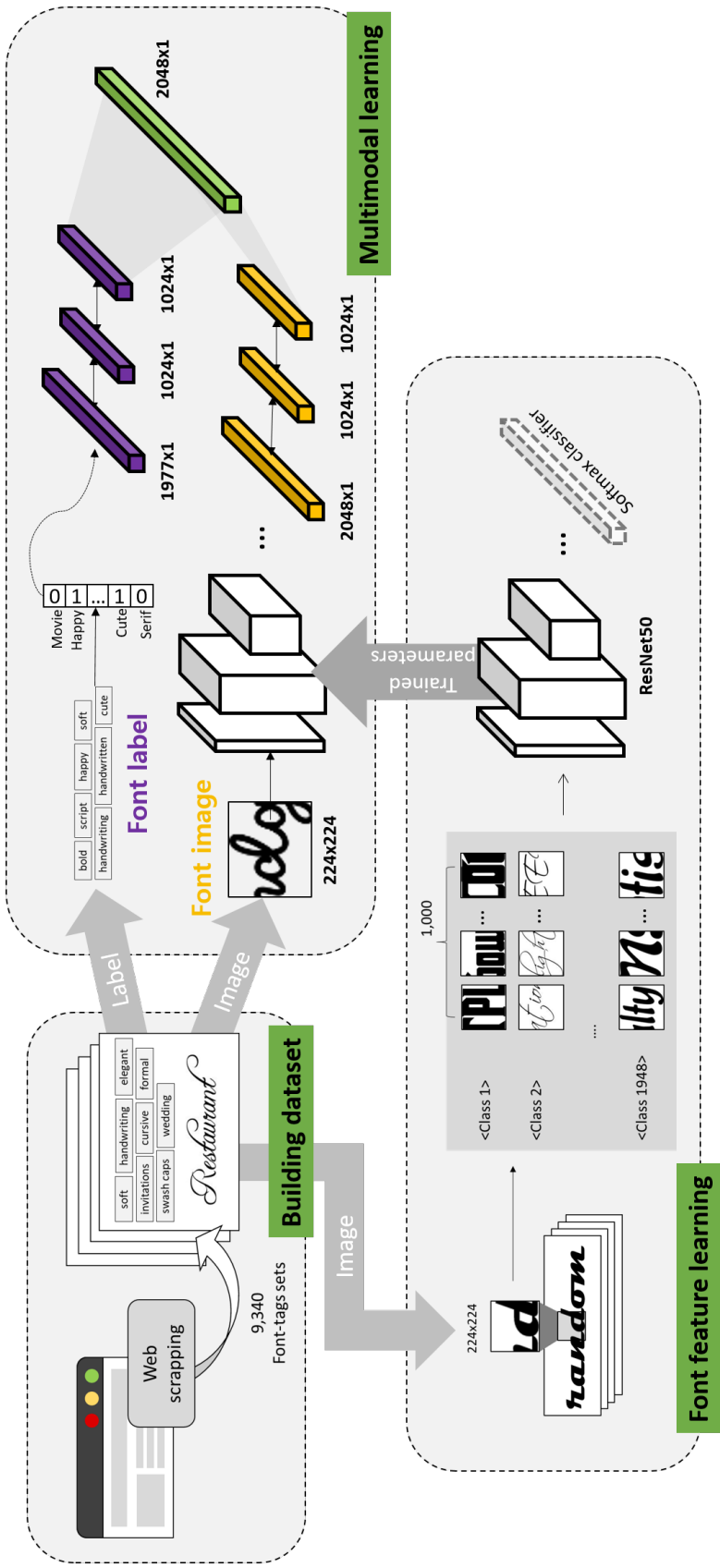


Fig. 7.2 The proposed framework is divided into three parts: building the dataset, font-feature learning, and multimodal learning. We collected large sets of a font with the associated tags and then used them to learn the visual features of the fonts and associated text.

to this, the motivation of the present study originated from a more human–computer interaction perspective: “How can machines be of assistance in realizing human creativity?”

The main contributions of this chapter are as follows:

- We propose a font search framework that uses an unsupervised generative model based the multimodal learning method that jointly learns the visual features of fonts and the associated tags for the creative process.
- We collected a large set of fonts with the associated tags that were annotated by professional level artists in order to utilize expert knowledge for the font domain.
- We present how newly created tags obtained through our framework are unexpected but useful concepts that are relevant to original fonts’ visual characteristic via qualitative and quantitative study.
- We investigate whether the concepts obtained from the model contribute to human creativity by conducting word association game with 12 participants.

As we can find in the above, the novelty of this chapter is in proposing a new way of using unsupervised generative model rather than achieving technical improvement. In the remainder of this ch, we review related studies on creativity and font-related works, illustrate the framework of the proposed method, present the outputs of the proposed framework with various input modalities, and finally present a user study and the analysis of its results.

7.2 Related Works

7.2.1 Machines and creativity

One of the fascinating questions in the machine learning field is “can artificial intelligence be creative?” Contrary to the conventional wisdom that creativity is only a human intellectual trait, several studies have shown fascinating results with the use of machines with visual, auditory, and/or other application. Image Analogies [65] is a framework that transfers textures from a source image to a target image. Using convolutional neural networks (CNNs), Neural Style Transfer [52] separates styles from images and synthesizes an image using content-independent representations. In addition to transferring the visual style of images, several studies have developed modeling machines that can create an image that has not been observed in the dataset [73, 124], compose music [165], and even write poetry [163]. Although all the aforementioned works demonstrate machine creativity, it has been considered rarely how these technologies can raise human creativity. Several works, meanwhile, have concerned with how artificial intelligence can affect human intellectual ability such as memorability [77, 136]. Siarohin et al. proposed an approach that finds the best style filters that synthesize an image to be more memorable [136]. Some studies devised tools to improve users’ experience in the creative process. PortraitSketch [159] is a drawing assistance system that automatically adjusts user-drawn

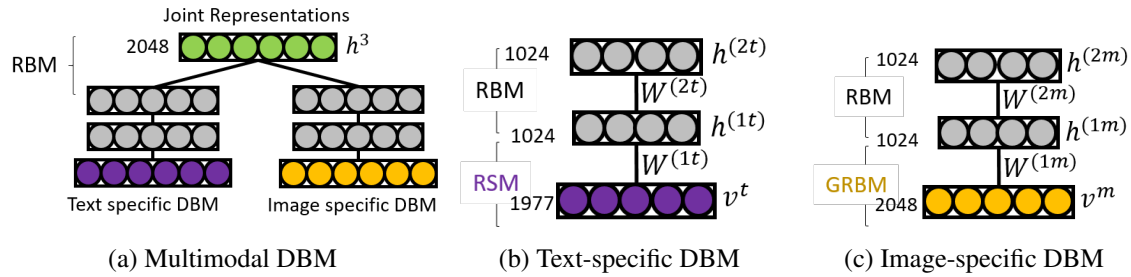


Fig. 7.3 A multimodal DBM (left) and two modality-specific DBMs (right two). The text-specific and image-specific DBMs are used for modeling sparse word count vectors and real-valued dense image features respectively. There is one additional layer (green) that combines the two modality-specific models on top of them.

strokes to source portrait image considering aesthetic quality, and DrawFromDrawings [103] enabled users to refer sketch images in a database to produce better drawings.

7.2.2 Affective effects of font

Font is a visual style of written language and it has increased and diversified over the century. As the font market grows, fonts are acknowledged as one of the most effective marketing tools for communicating the message [28, 63]. To describe the visual features of fonts, a typographic-specific dictionary has been developed [92]. For example, serif and sans serif are two of the most commonly used terms to describe the shape of the end of the stroke of a letter. The meaning of the words “bold” and “script” changes when we use these words in the typographic domain — these indicate the letters that are more weighted than others and those with a hand-drawn look, respectively. Several studies have been conducted in an attempt to determine the correlation between the typography-specific characteristics and emotional effect [135, 10] of fonts, but the obtained results are still conflicting and not well-deciphered yet because the visual appearance of fonts has numerous connotations. Nevertheless, the typographic terminologies are useful; several font web communities organize large font datasets using these terminologies [38, 2, Adobe]. Technological advancement has enabled people to access large font datasets in various ways. The visual font recognition (VFR) problem has thus been addressed. The initial large-scale VFR study achieved 72.50% of top 1 accuracy with 2,420 classes with local image descriptors (e.g., SIFT or LBP) [23]. Through the use of deep CNNs, the classification performance obtained has significantly improved so that it is applicable to a real-world image such as a signboard [154]. The exploratory search interface allows users who do not have any clear objective in mind to find fonts. Font Map [IDEO] is a new visualization tool that maps fonts in 3D space using visual similarity to allow users to explore in a new manner. One study presented an exploratory interface that organizes large font datasets using attributes such as “dramatic.” There exists a new type of font search system that recommends a font that harmoniously matches with a given user input image [31]. In contrast to the aforementioned tools designed for supporting efficient

search activity, this study investigates how the proposed font search framework affects the users' creativity and what they can obtain during the searching process with the system.

7.2.3 Multimodal deep boltzmann machine

To develop a model that learns the visual features of fonts and associated tags jointly, we make use of the restricted-Boltzmann-machines-based multimodal learning algorithm and multimodal deep Boltzmann machine (MDBM) [141]. The performance of MDBM has been measured in terms of accuracy in a classification task by learning fused representation jointly. In addition to this work, other multimodal setting studies [69] have taken into consideration the main contribution of the use of heterogeneous data in improving the accuracy. Although the aforementioned MDBM paper presented various examples that involved the filling-in of the missing data, they limited the capacity of the model to retrieval or classification tasks. Therefore, based on the idea that MDBM is a good generative model for filling-in a missing modality, we would like to take advantage of this model as a tool for generating inspiration that provides users unexpected but useful concepts. The more explanation of the model is described in the Section 3.2.

7.2.4 Inspiration

Memory and analogy are the fundamental prerequisites of the creative process [126, 108]. According to Guilford [57], creativity is related to divergent thinking — the ability to suggest various possible solution for a given problem. In addition, the suggested possibilities should not be random but should be related to the given problem. Therefore, to develop a new and good idea, it is important to have not only sufficient prior knowledge regarding the problem but also the ability to infer a link between the problem and the learned knowledge.

Our proposed framework reflects the above studies (see Figure 7.2). The building dataset part in Figure 7.2 is similar to the memory, collecting high-quality data, and building the prior knowledge processes. The font-feature learning and multimodal learning parts involve the learning of the visual features of the fonts and associated text information; these two parts are similar to the analogy process — solving problems by finding a relationship between observations. In the following section, we explain the proposed framework in detail.

7.3 Methodologies

In this section, we propose an approach that utilizes a large amount of expert knowledge that is accessible on the web. After a model learns the data, it is expected that it will generate associated information that will inspire users. Figure 7.2 shows the framework of the proposed method, and it is divided into three parts. In the following subsections, we explain each of the aforementioned parts. First, we describe how we collected the data for the dataset and then provide detailed descriptions of its contents. We then describe a model that jointly learns the visual features of the fonts and associated

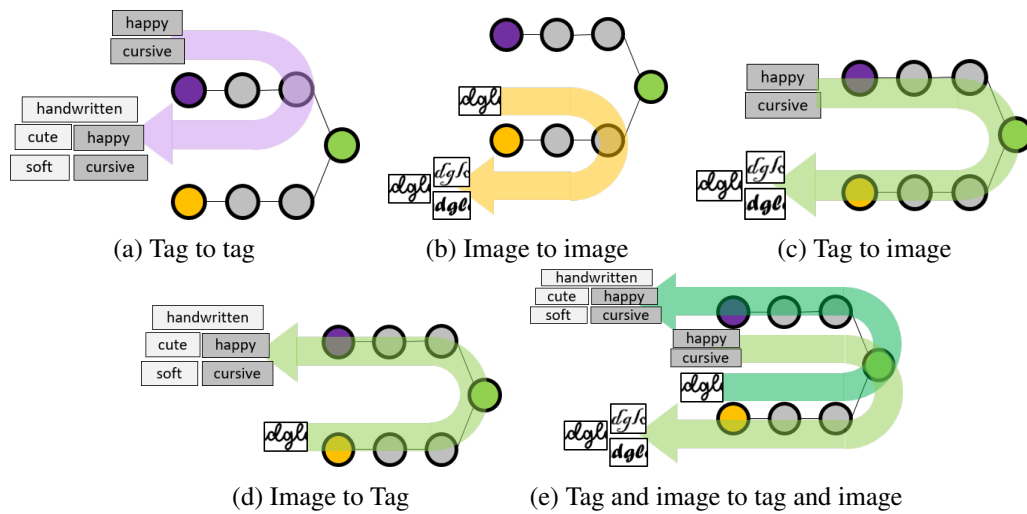


Fig. 7.4 Five path variations across the model. Given the text input (tag), the model can reconstruct the tag conditioned on the input (a). Similarly, if we input the image, the model retrieves images that share a visual characteristic that is similar to the given input image. On passing through the joint layer, the model outputs a different modality from the input (c, d, and e).

tags, and then demonstrate how this model can be used as a tool for inspiration. We then present the architecture of our model that extracts visual features of font images that will be used to feed the multimodal model.

7.3.1 Dataset

There is a font dataset [154], which is the largest and a recent dataset (2383 unique fonts). It provides a list of fonts but does not include the font file owing to copyright issues. Furthermore, it does not provide any associated tag. Thus, we use newly collected fonts and associated tags from the font sharing web communities. There are several font communities¹, and over 169,000 fonts exist in only one font community. Professional and amateur designers usually share their fonts through these communities, and the fonts are properly annotated with tags. We scraped web pages from the font web communities to collect fonts that have tags. We collected fonts with associated tags, and the number of font–tag pairs obtained reached 9,340. There are 3,715 unique tags across the dataset, and each font was annotated with an average of 8.23 tags.

7.3.2 Multimodal deep boltzmann machine

To learn the visual features of the fonts and associated tags jointly, we make use of the RBM-based multimodal learning algorithm, MDBM [141]. MDBM learns correlations across various modalities, which enables the sampling of a missing modality from the probabilistic model when one of the modalities is missing. For example, given a text, it samples images that share the same concept as the

¹Dafont: <https://www.dafont.com/>, 1001Fonts: <http://www.1001fonts.com/>, and Myfonts: <https://www.myfonts.com/>

given text input $P(v_{img} | v_{text})$ and vice versa. This can be used for information retrieval. In addition, it learns the correlation within a single modality. For the given tag, the model outputs the probability of each tag to be observed together $P(v_{text1} | v_{text2})$. For example, for the given tag “elegant,” the model outputs a high probability of the word “stylish.” This enables us to handle noisy data (especially tags). We use this ability to fill-in the missing modality by expanding the input data through learned association.

In the following subsections, we explain RBM, which is a neural network, and generalized models of RBM that can be used for modeling diverse modalities such as text and images. We then describe the advantage of using this model as a tool for inspiration.

Restricted Boltzmann Machines

is an undirected, probabilistic generative graphical model with stochastic binary visible units (visible layer) and stochastic binary hidden units (hidden layer). Every visible and hidden unit is connected to each other, but there is no connection between two nodes of the same layer. RBM defines an energy function as

$$E(v, h; \theta) = - \sum_{i=1}^D \sum_{j=1}^F W_{ij} v_i h_j - \sum_{i=1}^D b_i v_i - \sum_{j=1}^F a_j h_j, \quad (7.1)$$

where $\theta = \{W, a, b\}$ is the set of parameters. W_{ij} represents the interaction term between the visible unit v_i and hidden unit h_j ; and b_i and a_i are bias terms of the visible and hidden units, respectively.

Gaussian RBM

has been proposed for modeling real-valued inputs for vision and speech tasks [30]. The energy function is similar to the RBM, but Gaussian RBM differs in that it assumes that the input data (i.e., the visible units) is not binary, but follows a Gaussian distribution. This is defined as

$$E(v, h; \theta) = - \sum_{i=1}^D \sum_{j=1}^F \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{i=1}^D \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^F a_j h_j, \quad (7.2)$$

where σ_i is the standard deviation, and μ_i is the mean of the distribution of the visible unit v_i .

Replicated Softmax Model (RSM)

is another extension of the RBM to model sparse count data such as words in a document [68]. In a manner similar to the RBM, each visible unit has a binary value that indicates whether a word exists in the input document. Let us assume that K is the vocabulary size of the dataset to be learned, and the input document can be represented as a vector of the visible units, the length of which is K in the bag-of-words model. The energy function of RSM is defined as

$$E(v, h; \theta) = - \sum_{k=1}^K \sum_{j=1}^F W_{kj} v_k h_j - \sum_{k=1}^K b_k v_k - M \sum_{j=1}^F a_j h_j, \quad (7.3)$$

where M is the total number of words observed in a document. As [141] stated, if the size of the tags of each document does not exhibit a large variance, it can be simplified by fixing M as 1. Similarly, we also fixed the value of M as 1.

For every given energy function for each RBM, the joint distribution over the visible and hidden units is defined as

$$P(v, h; \theta) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta)), \quad (7.4)$$

where $Z(\theta)$ is a normalization constant.

Multimodal Deep Boltzmann Machines

We now introduce a model that jointly learns two different modalities. Figure 7.3 shows an MDBM that comprised an image-specific DBM, a text-specific DBM, and one additional layer that combines the two models on top of them. Each modality-specific DBM is built using multiple RBMs stacked on top of each another. The image-specific DBMs (Figure 7.3c) are built with two RBMs: the Gaussian–binary RBM and binary–binary RBM. The probability of v^m that is assigned by image-specific DBM given parameters θ^m is defined as

$$\begin{aligned} P(\mathbf{v}^m; \theta^m) &= \sum_{h^{(1m)}, h^{(2m)}} P(v^m, h^{(2m)}, h^{(1m)}; \theta^m) \\ &= \frac{1}{Z(\theta)} \sum_{h^{(1m)}, h^{(2m)}} \exp\left(-\sum_{i=1}^D \frac{(v_i^m - b_i^m)^2}{2\sigma_i^2} + \sum_{i=1}^D \sum_{j=1}^{F_1} W_{ij}^{(1m)} \frac{v_i^m}{\sigma_i} h_j^{(1m)} \right. \\ &\quad \left. + \sum_{j=1}^{F_1} \sum_{l=1}^{F_2} W_{jl}^{(2m)} h_j^{(1m)} h_l^{(2m)} + \sum_{j=1}^{F_1} b_j^{(1m)} h_j^{(1m)} + \sum_{l=1}^{F_2} b_l^{(2m)} h_l^{(2m)}\right), \end{aligned} \quad (7.5)$$

where D , F_1 and F_2 are the number of visible units, first hidden units and second hidden units respectively. $\theta^m = \{W^{(1m)}, W^{(2m)}, b^m, b^{(1m)}, b^{(2m)}\}$ is the set of interaction terms and bias terms. The text-specific DBM (Figure 7.3c) is built with a replicated Softmax model and binary–RBM. Similarly, the probability that the text-specific DBM assigns to a visible vector v^t is $P(\mathbf{v}^t; \theta^t) = \sum_{h^{(1t)}, h^{(2t)}} P(v^t, h^{(2t)}, h^{(1t)}; \theta^t)$. Third layer that combines the two models on top of two modality-specific DBMs is illustrated in Figure 7.3a, and it is defined as:

$$\begin{aligned} P(\mathbf{v}^m, \mathbf{v}^t; \theta) &= \underbrace{\sum_{h^{(2m)}, h^{(2t)}, h^{(3)}} P(h^{(2m)}, h^{(2t)}, h^{(3)})}_{\text{Layer for joint representation}} \\ &\quad \underbrace{\sum_{h^{(1m)}} P(v^m, h^{(1m)}, h^{(2m)})}_{\text{Image-specific term}} \underbrace{\sum_{h^{(1t)}} P(v^t, h^{(1t)}, h^{(2t)})}_{\text{Text-specific term}}. \end{aligned} \quad (7.6)$$

Because exact maximum likelihood learning is intractable to perform exact maximum likelihood, contrastive divergence—an approximated learning method—is generally used for training the MDBM.

Here, we briefly describe how the approximation works. First, we sample all the hidden units conditioned on the observation $P(\mathbf{h} | v; \theta)$, and we then reconstruct the visible units by sampling from the sampled hidden value $P(\mathbf{v} | h; \theta)$. This reconstructed visible vector is used to optimize the model. In the case of the binary–RBM, we can write the reconstruction process using Eqs.(1) and (4) as:

$$\begin{aligned} P(h_j = 1 | v) &= \text{sigmoid}\left(\sum_D^{i=1} W_{ij}v_i + a_j\right), \\ P(v_i = 1 | h) &= \text{sigmoid}\left(\sum_F^{j=1} W_{ij}h_j + b_i\right). \end{aligned} \tag{7.7}$$

You can find more details and inferences at this reference [141].

Advantage of using MDBM

We can expect the following advantages of using MDBM as a tool for inspiration. Because MDBM can handle noisy input data, it is useful with our collected font dataset as the associated tags applied to a given font are subjective. In other words, the MDBM, which is also used for topic modeling in which the correlation between words is learned, allows incomplete font label sets to complement each other. Let us assume that an expert has labeled a font as “futuristic.” In this case, he/she did not really mean that the font is “futuristic,” but rather “futuristic-like.” Let us also assume that the expert labeled a different word with a similar meaning as “science-fiction.” Thus, two different words that are similar to each other may be classified using the same topic such that the two words can complement each other.

We can learn from the above example that MDBM reflects the notion of inspiration — “the inspired ideas and associated information are combined together” [3]. Therefore, we propose the use of the MDBM to obtain inspiration through five path variations across the model (Figure 7.4). We can use a set of tags, a font image, or both of them as an input. On following the pathway depending on the input modality, we can observe the reconstructed tags and/or font image conditioned on the input. Here, we believe that given a single tag, an expanded tag set is unexpected but useful (Figure 7.4a). We expect that it can be used as a source for generating with real font-design applications. We present a detailed example of this in Section 4.

7.3.3 Font visual features

Instead of considering each pixel of the font image as a visible unit; we used visual features as an input for the image-specific DBM. Motivated by the success of CNNs in large-scale visual discriminative tasks, we developed a VFR architecture based on ResNet50 [61] for a new set of font classes and then extracted font features using the font recognition model.

Table 7.1 Examples of reconstruction given a single word and on following the pathway illustrated in Figure 7.4a.

Input	Reconstructed tags
Elegant	Elegant, Fashion, Graceful, Hairline , Luxury, Classic, Stylish, Cursive , Pretty, Fashionable, Wedding, Invitations, Glamorous, Feminine
Cute	Cute, Trendy, Smiley, Stars, Princess, Adorable, Happy, Faces, Romantic, Animals, Teen, Girly, Bees
Messy	Messy, Handwritten , Scribbled , Cramped, Mixed Case , Scrawled, Handwriting, Ugly, Mixed Up , Helvetica, Blotchy, Masculine, Staggered, Filled
BlackLetter	Blackletter, Old German, Calligraphy , Gothic, 1900s, Schwabacher, Fraktur, Textura, Old English, Gothic, Minuscule, Rotunda, Fountain, Pen , 1800s

Dataset

As described in Section 3.1, we collected 9,340 unique fonts from the font sharing web community. Because it is inefficient to define a classification problem using 9,340 classes, we considered a subset of the entire dataset that comprised 1,948 unique fonts out of 9,340 fonts. The classification of 1,984 fonts is the task for which the CNN is trained. In our setting, we considered English fonts only.

We prepared a large-scale font dataset using a previous font-recognition study setup [23] (Figure 7.2). We selected 3,847 words comprising more than four characters from 5,000 frequently used English words. We then used these words as a seed word corpus for rendering font images. For a given font, we randomly select 1,000 words from the corpus and then rendered 1,000 different font images. Because the size of the images varies with the font, we cropped the font image to the size 224×224 . Similarly, we generated 100 font images (10% of training images) for each font category for training. As a result, we produced 1,948,000 font images for the training and 194,800 images for the test.

Model Learning and Font Features

We trained ResNet50 from scratch with the large synthesized font-image dataset. We used cross entropy as the loss function, and the batch size was set as 64. It was implemented using Tensorflow background Keras and run on a server with Intel Xeon 2.60 GHz and four Tesla K80 GPUs. It took around 21 hours to complete the entire training. Using this trained model, we obtained a 96.9% accuracy with the test set. To extract the visual features for the image-specific DBM, we used the output before the last Softmax layer, which comprises 2,048D dense representations that will be fed to the Gaussian RBM.

Table 7.2 The top five images obtained given an image input and on following the pathway illustrated in Figure 7.4b. All retrieved images look similar but are rendered by different font file.


























Input	Retrieved images
	    
	    
	    
	    

Table 7.3 Example of filling-in missing modality (text) given a single modality (image) on following the pathway illustrated in Figure 7.4d.

Input	Inferred tags
	Script, Cursive , Handwritten, Handwriting, Swash Caps , Wedding, Invitations , Formal , Elegant , Long Ascenders , Beautiful , Upright ,
	Comic, Outlined, Rounded , Bold Outlines , Bubbly , Puffy , Shaky , Thick Outlines , Pimpled , Cramped , Chubby , Polkadot , Cheese , Komika
	Serif, Bracket Serif , Classic , Antique , Roman , Times New Roman , Venetian , Times , Clarendon , Garamond , Caslon , 1500s , Theano , Family

7.3.4 Model architecture for font image and tag MDBM

As can be observed in Figure 7.2, the model is divided into three parts. For the image-specific part, a 2,048D feature vector extracted from the trained ResNet50 is fed to a Gaussian RBM with 2,048 visible units and 1,024 binary hidden units, and the output is connected to another RBM with 1,024 and 1,024 binary visible and binary hidden units, respectively. For the text-specific part, a word vector size of 1,977 is fed to an RSM with 1,977 binary visible units and 1,024 binary hidden units, and the output is connected to another RBM with 1,024 and 1,024 binary visible and binary hidden units, respectively. One additional RBM with 2,048 and 1,024 binary visible and binary hidden units, respectively that combines the two models on top of them comprises the top-layer.

We pretrain the modality-specific deep networks in a greedy layer-wise unsupervised manner and then fine-tune the entire network with the learning framework proposed in [138].

7.4 Qualitative Analysis

As mentioned in the introduction, we propose the use of an MDBM to obtain inspiration through the five path variations across the model (Figure 7.4). A previous MDBM study [141] showed examples of a one-step reconstruction (e.g., sample all the hidden units conditioned on the given tag $P(h | v_{tag}; \theta)$, and the reconstruct the tag by sampling from the sampled hidden value $P(v_{tag} | h; \theta)$), however, the examples provided were very limited, and the study lacked a comprehensive analysis. In this section, we examine the reconstruction results through all the possible data pathways across the architecture (Figure 7.3) and then illustrate the promising interventions that can be obtained on using the model.

7.4.1 Tag to tag

Table 7.1 lists examples of the reconstruction given a single word and on following the pathway illustrated in Figure 7.4a. It can be observed that the reconstructed tags obtained on using the model share the same concept as the given input word (e.g., “graceful” or “luxury” for a given word “elegant”). A more interesting observation is that we could observe how the writing style (e.g., “handwritten”) or typographical characteristics (e.g., “hairline,” which means a thin stroke) could be related to the impression of fonts. We can say that the hairline or cursive type fonts create an elegant feeling. Using the “mixed cased” font, a typographical feature that mixes the uppercase and lowercase letters gives a sense of “messy.” Furthermore, we can obtain information regarding the origin of the font — the unique stroke style of Blackletter is one of the characteristics of calligraphy written with a pen named nip, which was used throughout mid-age western Europe.

7.4.2 Image to image

Table 7.2 lists the top five images obtained given an image input and on following the pathway illustrated in Figure 7.4b. It can be observed that the system properly retrieved the query font images

	Elegant	Cute	Messy	Halloween
Elegant				
Cute				
Messy	Random images with low probability			
Halloween		Random images with low probability		

Fig. 7.5 Example of retrieved images of a given single topic (diagonal) and two tags from different topics (lower triangular) on following the pathway illustrated in Figure 7.4c.

for the given query as the top-1 font. Furthermore, the retrieved fonts have a very similar appearance but are actually different from each input font.

7.4.3 Image to tag

Table 7.3 lists examples of filling-in missing modality (text) given a single modality (image) and on following the pathway illustrated in Figure 7.4d. The black tags indicate the original tags of the given input image, and the blue tags are newly associated tags. Let us look at the new tags for the first script font. We can find not only the impressions the font creates (e.g., elegant or beautiful) but also the application of the font (e.g., invitations). These fonts would remind us of the cover of wedding invitations. The expanded tags for the second font appear to be more diversified. Rounded fonts with bold outlines give a sense of touch such as bubbly and puffy. The tag “cheese” may remind one of camembert cheese with a hole. The newly associated tags for the third font have much information that is related to the origin of the font (e.g., Roman, 1500s).

7.4.4 Tag to image

Figure 7.5 shows the retrieval results of a given single topic (diagonal) and two tags from different topics (lower triangle). For example, the fonts retrieved by a single input “elegant” are cursive or have light strokes that appear to be written softly and carefully. Let us observe another example — the

Table 7.4 Example of multimodal output from multimodal input on following the pathway illustrated in Figure 7.4e.

Input image and tag	Retrieved images and reconstructed tags
 <p>Halloween</p>	 <p>Halloween, Horror, Scary, Creepy, Bloody, Vampire, Sinister, Eerie, Witches, Zombie</p>
 <p>Elegant</p>	 <p>Elegant, Wedding, Hairline, Long Ascenders, Feminine, Party, Spiral,</p>
 <p>Messy</p>	 <p>Messy, Mixed Up, Brushed, Typewriter, Mixed Case, Gangster, Blotchy</p>
 <p>Blackletter</p>	 <p>Blackletter, Calligraphy, Old German, Gothic, Old English, Fraktur, 1900s,</p>

elegant–cute combination. None of the fonts include either elegant or cute as their tags. Nonetheless, all the fonts look elegant and cute — the fonts have a detail at the end of their stroke that gives a lovely and cute impression to the font.

In the case of the cute–messy combination, it can be observed that all the fonts look like they have been written playfully and roughly. Interestingly, there is only one font that has the tags *cute* and *messy* as their original tags (red bounding box). Although the fonts did not contain the word *cute* or *messy* as their original tag, the model that learned the correlation between the visual features and the tag responded to the tag input appropriately.

Let us consider another example. It is difficult to remind the concept of *messy* and *Halloween* from the following tags, *Outlined*, *Grunge*, *Stamped*, *Narrow*, *Crooked*, *Bold*, *Display*, *Black & White*, and *Letterpress*. This is the tag set for the blue-bordered font in the *Halloween–messy* cell, and the font fits very well with its topic. However, the tags are difficult for us to imagine. From these observations, we can expect that although the original tag is sparse and has low relevance, the model provides useful results by jointly learning the visual features of the font and associated tags. However, if the two concepts are contrary to each other, the results do not seem to be successful— 19 fonts have both “cute” and “Halloween” as their tags, but the model outputs random results.

7.4.5 Tag and image to tag and image

We also obtain a multimodal output from the multimodal input on following the pathway illustrated in Figure 7.4e. Table 7.4 illustrates examples of multimodal output from the multimodal input. Here, we can see top retrieved results are visually similar images to the given image input, and semantically analogous tags to the given text input. The main contribution of the previous MDBM study is that a better retrieval result is obtained by learning multimodal features in terms of mAP measure. We can see Table 7.4 shows this tendency. In addition to robustness, we propose a variety use of MDBM — helping users to come up with diverse concepts.

As stated in the introduction, tags for font comprise heterogeneous subjective information such as impressions (e.g., *elegant*), applications (e.g., *invitation*), and typographic-specific features (e.g., *handwritten*). Owing to this nature of the font dataset, we were able to observe probable, but unexpected tags that would help users expand their idea of the given font. Figure 7.6b shows examples of expanded tags to the given multimodal input. Although some fonts share a specific visual feature, every font has a unique appearance of its own. Figure 7.6a shows four visually different fonts that have the same typographical feature of *low contrast* — fewer changes in the thickness of the stroke. We anchored the tag input to *Low contrast*, and then obtained the different tag sets for the varying font image input. We thus expect to know what concepts were assigned to each font by the typographical feature *Low contrast*.

Although it is difficult to determine what impression the font creates or where it can be used from the original tags, the multimodal path produced interesting concepts such as impression (e.g., *stamped*, *wood*) and/or its application (e.g., *sports*, *brandname*). We highlighted several tags that are probable,



Fig. 7.6 Expanded tags from anchored tag and varying font image input following the pathway illustrated in Figure 4(e).

but unexpected tags that would help users expand their idea of the given font in blue. The tags *wood*, *stamped*, and *soft* for the font in top left provide a new perspective that the font seems to match well with the inscription on the warm feeling wood. The tags in top right remind one of a notebook that generates impressions of neatness and tidiness. In the case of the bottom left font, tags that reflect the impression of a sports brand that is old-fashioned are ranked at the top (e.g., 1960s, sports, and brand). This font is similar to the logo of a sports brand that is noted for its vintage style. In case of the bottom right font, we might be not able to imagine the tags *stars* and *flower* from the original tags. In addition, if handwritten-type fonts are given, the system specified the tool that was used for writing the font such as nip pen, marker pen, or brush (See Figure 7.6b). As results showed, the model suggested interesting concepts that helped us to jump from the given topic to another. In the next section, we investigate whether the suggested concepts contribute to human creativity via a user test.

7.4.6 User test

To measure how much the system helps in enhancing the human creative process, we used a *word-association game*, which is a known metric for measuring creativity [16]. The word-association game allows a participant to write down words that come to their mind about a given problem. We devised a question to measure how well a user comes up with different solutions to a single problem.

Table 7.5 Retrieval performance (*Recall@K*) comparison of original tag set ORI and randomly removed original tag set ORI-1, ORI-2 with generated tag set GEN, GEN-1, and GEN-2. #Tags indicates the average number of tags in a tag set.

<i>K</i>	CP	ORI	GEN	ORI-1	GEN-1	ORI-2	GEN-2
10	0.001	0.305	0.241	0.200	0.215	0.136	0.189
20	0.002	0.402	0.333	0.274	0.306	0.194	0.274
30	0.003	0.465	0.396	0.325	0.367	0.235	0.333
#Tags		7		6		5	

Task and participants

After selecting 15 fonts that have a different appearance from those in the dataset, we prepared 15 pairs of font with originally associated tags. We also prepared 15 pairs of fonts with expanded tags by the model. This allows participants to perform two tasks with different tag sets for each font. Here, the number of tags for each font has been adjusted to the number of original tags, so that we restrict the amount of concepts exposed to participants. The total number of pairs is 30. We randomly showed the 30 font–tag pairs and requested participants to write down as many associated concepts that come into their mind as possible in 30 s. Because of the diversity of participants’ nationalities, we presented the translated tags in their own language and the participants were requested to write in their native language. We provided each participant a laptop for writing. ten questions constituted one session, and after one session, the participants were given 60 seconds break. A total of 12 people participated, and every participant answered all the questions. We totally collected 360 answers.

Results

The number of concepts that users wrote down is referred to as the idea fluency. We compared the idea fluency between the original tags and expanded tags, and a paired-samples t-test was conducted on the collected 360 answers (180 pairs). The participants showed better idea fluency with the expanded tags (mean=4.34, sd=3.70) than the original tags (mean=4.12, sd=2.89); significant at the .05 level ($p = 0.04$). This result demonstrates the idea that the model can be used as a tool for inspiration.

7.5 Quantitative Experiment

From the qualitative analysis above, we showed that expanded tags obtained through our framework are unexpected but useful concepts that seem to be relevant to original fonts’ visual characteristics. Assuming the system generates unexpected concepts, we evaluate how much the concepts are useful by providing font retrieval performance.

Table 7.6 Retrieval performance ($Recall@K$) comparison of adding random generated tags (RAN5/RAN10) with adding system generated tags (GEN5/GEN10).

K	ORI+RAN5	ORI+GEN5	ORI+RAN10	ORI+GEN10
10	0.252	0.273	0.173	0.247
20	0.340	0.367	0.243	0.343
30	0.400	0.431	0.292	0.408
#Tags	12		17	

7.5.1 Evaluation

Let us assume we have a font i and an associated tag set T_i . To evaluate the performance of the tag set T_i to retrieve the font i , we get ranked retrieval results given the tag set T_i through the tag to image path of the MDBM, and then calculate $Recall@K$ over whole dataset. Given font i , we compare $Recall@K$ between original tag set ORI_i , generated tag set GEN_i , damaged tag sets of each, and noisy tag sets of each over whole dataset. Here, tag set ORI is the original tag set, which can be regarded as ground truth. By comparing the retrieval performance, we would like to show the generated tags by given font image i are not only new (unexpected), but also relevant to the font i . We compare the performance of font sets that have different characteristics as follows.

Damaged tag sets: The tag sets ORI-1 and ORI-2 are randomly deleted tag sets by one and two from the tag set ORI respectively. We call the randomly deleted tag set as damaged tag set. Tag set GEN-1, and GEN-2 are generated tag sets, and the each number of tags to generate were set to the same number as ORI-1, and ORI-2 respectively.

Noisy tag sets: Noisy tag sets indicate the tag sets that contains additional tags that are not originally annotated. The tag set ORI+RAN5 and ORI+RAN10 are noisy tag sets that have additional random five or ten tags respectively. On the other hand, we added five or ten system generated tags, ORI+GEN5 and ORI+GEN10.

7.5.2 Results

Table 7.5 shows $Recall@K$ depending on damaged tag sets. The column CP indicates Chance Performance. As we can see in the Table 7.5, the smaller the number of tags is, the lower the retrieval performance is (ORI > ORI-1 > ORI-2, GEN > GEN-1 > GEN-2). Original tag set (ORI) showed the best performance. Nevertheless, in the case of damaged tag sets, generated tag sets produced better performance than original tag sets (GEN-1 > ORI-1, GEN-2 > ORI-2). Table 7.6 shows $Recall@K$ depending on noisy tag sets. We can find that the generated tag sets show better retrieval performance than the noisy random tag sets.

By integrating our qualitative analysis with the quantitative results, we can say that the generated concepts are not only new but also relevant to each font and the proposed framework can be used as a tool for inspiration that helps users to realize what they imagine by expanding their idea.

7.6 Conclusions

We proposed a framework that helps users to explore a font dataset using the multimodal method and provides unexpected but useful font images or concepts for supporting the creative process. We collected a large number of fonts and their associated tags in a dataset. We proposed a new way of using unsupervised generative model MDBM for a tool for generating inspiration. By examining the newly associated tags and retrieved images, we observed unexpected but useful results. The model not only suggested interesting tags but also reconstructed the tags that reflect the fonts' characteristics such as which tool is used for writing the font. In the user study, we observed that the participants produced more diverse concepts when inspired by the generated tags. Based on these results, the model was found to be worthy of being used as a tool of association.

Chapter 8

Conclusion

8.1 Summary

In this thesis, we presented methods to understand affects in image and font, and proposed applications that the affect information can be useful. We summarize the major contributions of this thesis as follow:

Image impression retrieval: Conventional image retrieval systems ask users to input query by text. However, it is not always easy for users to convert their intention into verbal representations. In Chapter 3, we proposed an interactive retrieval system based on yes-no questions for image impression retrieval. We modeled a system that interprets images with impression words such as fresh and modern. Then, we introduced a yes-no question based querying method and a feedback interface to support users querying. From the user test, we showed that our system brings satisfactory results to users in case where the proper text querying is difficult.

Font emotion understanding: Different fonts create different experience. This ability of font have been utilized and studied in marketing and branding strategies (e.g., powerfulness of logo). However, there are few studies about font as an emotional modulator. In Chapter 4, we demonstrated the effect of fonts on viewer's emotional state by two experimental studies — explicit and implicit testing. In explicit testing, we measured the response to fonts using the assessment sheet which asks readers the feeling in the font directly. In implicit testing, we measured unconscious response to fonts using spontaneous speeches that elicited by fonts of written text. The series of studies showed the potential use of font for emotional representation such as happiness and anger.

Font communication on mobile messenger: Instant messaging is a popular form of text-based communication. However, text-based messaging lacks the ability to communicate nonverbal information such as that conveyed through facial expressions and voice tones. In Chapter 5, we proposed Emotype, a mobile messenger application prototype that enables users to change the font of a mobile messenger message to convey certain emotions. In user test, we demonstrated the feasibility of fonts for communicating emotions with a survey study, and then explored the unique feature of font that is different from other ways for expressing emotion by qualitative user study.

Font search by image: One of the important aspects in graphic design is choosing the font of the caption that matches aesthetically the associated image. In Chapter 6, we presented two font search systems that enable users to use images as queries - (1) query by image impressions based on color study and (2) query by image contents based on concept analysis. Instead of matching font and image directly, we mapped both image and font to color-based semantic space or concept-based semantic space. Our evaluation results showed that the recommended fonts scored better than other comparisons and provides competing results with the ones chosen by experienced graphic designers.

Creativity support in graphic design: Inspiration plays an important role in the creative process. By getting inspired, we can reach unexpected but useful ideas. Inspiration, generally, comes to us when we interact with external interventions. In Chapter 7, we presented a framework that assist users' interactions in font search with unexpected but useful concepts generated by multimodal learning. By examining the results of the model that change with various inputs, we observed that the model produces promising results that appeared to be useful for inspiring users.

8.2 Future Directions

We overview some possible future directions of this thesis. In Chapter 4, we demonstrate the effect of fonts by explicit and implicit testing. In addition to that, we expect to collect synchronized response to font by introducing physiological method such as functional magnetic resonance imaging (fMRI) and Electroencephalography (EEG). In Chapter 5, we enabled users to change the font of message. Here, the number of available fonts were very limited. In future, we expect to see more diverse experience among users by enlarging available fonts to use.

References

- [1] (2017). Esl fast. <http://www.eslfast.com/easydialogs/>. Accessed: 2017-06-28.
- [2] (2018). Fonts in use. <https://fontsinuse.com/>.
- [3] A. Jacko, J. (2009). *Human-Computer Interaction. Novel Interaction Methods and Techniques*, volume 5611.
- [4] Abelin, Å. and Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.
- [Adobe] Adobe. Typekit. <https://typekit.com/fonts>.
- [6] Alexander, J. D. and Nygaard, L. C. (2008). Reading voices and hearing text: talker-specific auditory imagery in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 34(2):446.
- [7] Ali, A. Z. M., Wahid, R., Samsudin, K., and Idris, M. Z. (2013). Reading on the computer screen: Does font type has effects on web text readability? *International Education Studies*, 6(3):26.
- [8] Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.
- [9] Amare, N. and Manning, A. (2012a). Seeing typeface personality: Emotional responses to form as tone. In *2012 IEEE International Professional Communication Conference*, pages 1–9. IEEE.
- [10] Amare, N. and Manning, A. (2012b). Seeing typeface personality: Emotional responses to form as tone. In *Professional Communication Conference (IPCC), 2012 IEEE International*, pages 1–9. IEEE.
- [11] Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4):245–266.
- [12] Arditi, A. and Cho, J. (2005). Serifs and font legibility. *Vision research*, 45(23):2926–2933.
- [13] Azadi, S., Fisher, M., Kim, V., Wang, Z., Shechtman, E., and Darrell, T. (2018). Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 11, page 13.
- [14] Baddeley, A. D. and Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, 8:47–89.
- [15] Baron, N. S. (2004). See you online gender issues in college student use of instant messaging. *Journal of Language and Social Psychology*, 23(4):397–423.

- [16] Benedek, M., Könen, T., and Neubauer, A. C. (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3):273.
- [17] Berkovsky, S., Taib, R., and Conway, D. (2017). How to recommend?: User trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 287–300. ACM.
- [18] Bernard, M. L., Chaparro, B. S., Mills, M. M., and Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed times new roman and arial text. *International Journal of Human-Computer Studies*, 59(6):823–835.
- [19] Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, pages 223–232.
- [20] Byron, K. and Baldrige, D. C. (2007). E-mail recipients’ impressions of senders’ likability the interactive effect of nonverbal cues and recipients’ personality. *Journal of Business Communication*, 44(2):137–160.
- [21] Campbell, N. D. and Kautz, J. (2014). Learning a manifold of fonts. *ACM Transactions on Graphics (TOG)*, 33(4):91.
- [22] Candello, H., Pinhanez, C., and Figueiredo, F. (2017). Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3476–3487. ACM.
- [23] Chen, G., Yang, J., Jin, H., Brandt, J., Shechtman, E., Agarwala, A., and Han, T. X. (2014a). Large-scale visual font recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3598–3605.
- [24] Chen, P., Fu, X., Teng, S., Lin, S., and Lu, J. (2014b). Research on micro-blog sentiment polarity classification based on svm. In *International Conference on Human Centered Computing*, pages 392–404. Springer.
- [25] Chen, T. (2017). Zi2zi: Master chinese calligraphy with conditional adversarial networks. <https://github.com/kaonashi-tyc/zi2zi>.
- [26] Chen, T., Borth, D., Darrell, T., and Chang, S.-F. (2014c). DeepSentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.
- [27] Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202. ACM.
- [28] Childers, T. L. and Jass, J. (2002a). All dressed up with something to say: Effects of typeface semantic associations on brand perceptions and consumer memory. *Journal of Consumer Psychology*, 12(2):93–106.
- [29] Childers, T. L. and Jass, J. (2002b). All dressed up with something to say: Effects of typeface semantic associations on brand perceptions and consumer memory. *Journal of Consumer Psychology*, 12(2):93–106.
- [30] Cho, K. H., Raiko, T., and Ilin, A. (2013). Gaussian-bernoulli deep boltzmann machine. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–7. IEEE.
- [31] Choi, S., Aizawa, K., and Sebe, N. (2018). Fontmatcher: Font image paring for harmonious digital graphic design. In *23rd International Conference on Intelligent User Interfaces*, pages 37–41. ACM.

- [32] Choi, S., Yamasaki, T., and Aizawa, K. (2015). An interactive system based on yes-no questions for affective image retrieval. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 45–50. ACM.
- [33] Choi, S., Yamasaki, T., and Aizawa, K. (2016). Typeface emotion analysis for communication on mobile messengers. In *Proceedings of the 1st International Workshop on Multimedia Alternate Realities*, pages 37–40. ACM.
- [34] Co., L. (2016). Line stickers. LINE STORE. <https://store.line.me/stickershop/showcase/top/en>.
- [35] Collins, N. L. and Miller, L. C. (1994). Self-disclosure and liking: a meta-analytic review. *Psychological bulletin*, 116(3):457.
- [36] Crawford, J. R. and Henry, J. D. (2004). The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265.
- [37] Cui, C., Shen, J., Ma, J., and Lian, T. (2017). Social tag relevance learning via ranking-oriented neighbor voting. *Multimedia Tools and Applications*, 76(6):8831–8857.
- [38] dafont.com (2016). Dafont. <http://www.dafont.com/>.
- [39] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5.
- [40] de Freitas, L. A., Vanin, A. A., Hogetop, D. N., Bochernitsan, M. N., and Vieira, R. (2014). Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 628–633. ACM.
- [41] Derks, D., Fischer, A. H., and Bos, A. E. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785.
- [42] Diao, Q., Qiu, M., Wu, C.-Y., Smola, A. J., Jiang, J., and Wang, C. (2014). Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM.
- [43] Diaz, F., Mitra, B., and Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 367–377.
- [44] Dror, O. E. (2001). Counting the affects: Discoursing in numbers. *Social research*, pages 357–378.
- [45] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.
- [46] Ekman, P. and Friesen, W. (1976). Pictures of facial affect (palo alto, ca: Consulting psychologists).
- [47] Fackrell, J., Vereecken, H., Buhmann, J., Martens, J.-P., and Coile, B. V. (2000). Prosodic variation with text type. In *Sixth International Conference on Spoken Language Processing*.
- [48] Filik, R. and Barber, E. (2011). Inner speech during silent reading reflects the reader’s regional accent. *PloS one*, 6(10):e25782.
- [49] Fleckenstein, K. S. (1991). Defining affect in relation to cognition: A response to susan mcLeod. *Journal of Advanced Composition*, pages 447–453.

- [50] Gao, X.-P. and Xin, J. H. (2006). Investigation of human's emotional responses on colors. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 31(5):411–417.
- [51] Gatt, A. and Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.
- [52] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2414–2423. IEEE.
- [Google] Google. Google cloud vision api. <https://vision.googleapis.com>.
- [54] Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2016). Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer.
- [55] Gottschalk, L. A. and Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.
- [56] Guest, G., Namey, E., and McKenna, K. (2017). How many focus groups are enough? building an evidence base for nonprobability sample sizes. *Field methods*, 29(1):3–22.
- [57] Guilford, J. P. (1977). *Way Behond the IQ: Guide to Improving Intelligence and Creativity*. Creative Education Foundation.
- [58] Gyurak, A., Gross, J. J., and Etkin, A. (2011). Explicit and implicit emotion regulation: a dual-process framework. *Cognition and Emotion*, 25(3):400–412.
- [59] Hancock, J. T., Landrigan, C., and Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932. ACM.
- [60] Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19.
- [61] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [62] Heery, M. W. (1989). Inner voice experiences: An exploratory study of thirty cases. *The Journal of Transpersonal Psychology*, 21(1):73.
- [63] Henderson, P. W., Giese, J. L., and Cote, J. A. (2004). Impression management using typeface design. *Journal of marketing*, 68(4):60–72.
- [64] Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM.
- [65] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM.

- [66] Himmelman, N. P. and Ladd, D. R. (2008). Prosodic description: An introduction for field-workers.
- [67] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.
- [68] Hinton, G. E. and Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614.
- [69] Huang, M., Rong, W., Arjannikov, T., Jiang, N., and Xiong, Z. (2016). Bi-modal deep boltzmann machine based musical emotion classification. In *International Conference on Artificial Neural Networks*, pages 199–207. Springer.
- [70] Huang, S., Zhong, Z., Jin, L., Zhang, S., and Wang, H. (2018). Dropregion training of inception font network for high-performance chinese font recognition. *Pattern Recognition*, 77:395–411.
- [71] Huang, Y., Wang, W., and Wang, L. (2017). Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318.
- [IDEO] IDEO. Font map. <http://fontmap.ideo.com/>.
- [73] Inc., G. (2014). Deep dream. <https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>.
- [74] Inc., G. (2016a). Allo. Apps. <https://play.google.com/store/apps/details?id=com.google.android.apps.fireball&hl=en>.
- [75] Inc., Y. (2016b). Yelp dataset. <https://www.yelp.com/dataset/challenge>.
- [76] Ishibashi, K. and Miyata, K. (2016). Edit-based font search. In *International Conference on Multimedia Modeling*, pages 550–561. Springer.
- [77] Isola, P., Xiao, J., Parikh, D., Torralba, A., and Oliva, A. (2014). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482.
- [78] Jahanian, A., Liu, J., Lin, Q., Tretter, D., O’Brien-Strain, E., Lee, S. C., Lyons, N., and Allebach, J. (2013). Recommendation system for automatic design of magazine covers. In *ACM IUI*, pages 95–106.
- [79] Josephson, S. (2008). Keeping your readers’ eyes on the screen: An eye-tracking study comparing sans serif and serif typefaces. *Visual communication quarterly*, 15(1-2):67–79.
- [80] Kalman, Y. M. and Gergle, D. (2014). Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior*, 34:187–193.
- [81] Kato, M. P., Yamamoto, T., Ohshima, H., and Tanaka, K. (2014). Cognitive search intents hidden behind queries: a user study on query formulations. In *WWW*, pages 313–314.
- [82] Kim, M., Choi, K., and Suk, H.-J. (2016). Yo!: Enriching emotional quality of single-button messengers through kinetic typography. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS ’16, pages 276–280, New York, NY, USA. ACM.

- [83] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer.
- [84] Kobayashi, S. (1981). The aim and method of the color image scale. *Color research & application*, 6(2):93–107.
- [85] Kobayashi, S. and Matsunaga, L. (1991). *Color image scale*. Kodansha international Tokyo.
- [86] Kosslyn, S. M. and Matt, A. M. (1977). If you speak slowly, do people read your prose slowly? person-particular speech recoding during reading. *Bulletin of the Psychonomic Society*.
- [87] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- [88] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- [89] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.
- [90] Kuzi, S., Shtok, A., and Kurland, O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1929–1932. ACM.
- [91] La Manna, S., Colia, A., and Sperduti, A. (1999). Optical font recognition for multi-font ocr and document processing. In *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, pages 549–553. IEEE.
- [92] Lawson, A. S. (1990). *Anatomy of a Typeface*. David R. Godine Publisher.
- [93] Layer (2016). Atlas. <https://atlas.layer.com/>.
- [94] Lee, J., Jun, S., Forlizzi, J., and Hudson, S. E. (2006). Using kinetic typography to convey emotion in text-based interpersonal communication. In *Proceedings of the 6th Conference on Designing Interactive Systems, DIS '06*, pages 41–49, New York, NY, USA. ACM.
- [95] Lee, S., De Neve, W., and Ro, Y. M. (2014). Visually weighted neighbor voting for image tag relevance learning. *Multimedia tools and applications*, 72(2):1363–1386.
- [96] Li, Y. and Suen, C. Y. (2010). Typeface personality traits and their design characteristics. In *proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 231–238. ACM.
- [97] Li, Z., Wu, X.-M., and Chang, S.-F. (2012). Segmentation using superpixels: A bipartite graph partitioning approach. In *CVPR*, pages 789–796.
- [98] Limited, C. G. (2016). Font infinity. iTunes. <https://itunes.apple.com/us/app/font-infinity-cool-new-fonts/id885791898?mt=8>.
- [99] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- [100] Loewenstein, G. (2005). Hot-cold empathy gaps and medical decision making. *Health Psychology*, 24(4S):S49.

- [101] Lopes, A. T., de Aguiar, E., De Souza, A. F., and Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628.
- [102] Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. pages 83–92.
- [103] Matsui, Y., Shiratori, T., and Aizawa, K. (2017). Drawfromdrawings: 2d drawing assistance via stroke interpolation with a sketch database. *IEEE transactions on visualization and computer graphics*, 23(7):1852–1862.
- [104] Mehrabian, A. (1980). Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies.
- [105] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- [106] Min, F. (2016). Font dresser. iTunes. <https://itunes.apple.com/us/app/font-dresser-for-ext-editing/id467286515?mt=8>.
- [107] Montefinese, M., Ambrosini, E., Fairfield, B., and Mammarella, N. (2014). The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.
- [108] Moorman, C. and Miner, A. S. (1997). The impact of organizational memory on new product performance and creativity. *Journal of marketing research*, pages 91–106.
- [109] Munezero, M. D., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.
- [110] NAz, K. and Epps, H. (2004). Relationship between color and emotion: A study of college students. *College Student J*, 38(3):396.
- [111] O’Donovan, P., Libeks, J., Agarwala, A., and Hertzmann, A. (2014). Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 33(4):92.
- [112] Ohene-Djan, J., Wright, J., and Combie-Smith, K. (2007). Emotional subtitles: A system and potential applications for deaf and hearing impaired people. In *CVHI*.
- [113] Ou, L.-C., Luo, M. R., Woodcock, A., and Wright, A. (2004). A study of colour emotion and colour preference. part i: Colour emotions for single colours. *Color Research & Application*, 29(3):232–240.
- [114] Oyama, T. (2003). Affective and symbolic meanings of color and form: Experimental psychological approaches. *Empirical Studies of the Arts*, 21(2):137–142.
- [115] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.
- [116] Park, T. W., Kim, S.-J., and Lee, G. (2014). A study of emoticon use in instant messaging from smartphone. In *International Conference on Human-Computer Interaction*, pages 155–165. Springer.

- [117] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- [118] Phan, H. Q., Fu, H., and Chan, A. B. (2015). Flexyfont: Learning transferring rules for flexible typeface synthesis. In *Computer Graphics Forum*, volume 34, pages 245–256. Wiley Online Library.
- [119] Picard, R. W. et al. (1995). Affective computing.
- [120] Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.
- [121] Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(03):715–734.
- [122] Pu, H.-T. (2008). An analysis of failed queries for web image retrieval. *Journal of Information Science*, 34(3):275–289.
- [123] Qian, X., Liu, X., Zheng, C., Du, Y., and Hou, X. (2013). Tagging photos using users’ vocabularies. *Neurocomputing*, 111:144–153.
- [124] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- [125] Rautaray, S. S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54.
- [126] Reeves, L. M. and Weisberg, R. W. (1993). On the concrete nature of human thinking: Content and context in analogical transfer. *Educational Psychology*, 13(3-4):245–258.
- [127] Rello, L., Pielot, M., and Marcos, M.-C. (2016). Make it big!: The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 3637–3648, New York, NY, USA. ACM.
- [128] Rubinstein, R. (1988). *Digital typography: an introduction to type and composition for computer system design*. Addison-Wesley Longman Publishing Co., Inc.
- [129] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.
- [130] San Pedro, J., Yeh, T., and Oliver, N. (2012). Leveraging user comments for aesthetic aware image search reranking. In *WWW*, pages 439–448.
- [131] Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6):2408–2412.
- [132] Schirmer, A. and Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in cognitive sciences*, 21(3):216–228.
- [133] Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81.
- [134] Schwartz, B. (2004). The paradox of choice.

- [135] Shaikh, A. D., Chaparro, B. S., and Fox, D. (2006). Perception of fonts: Perceived personality traits and uses. *Usability news*, 8(1):1–6.
- [136] Siarohin, A., Zen, G., Majtanovic, C., Alameda-Pineda, X., Ricci, E., and Sebe, N. (2017). How to make an image more memorable?: A deep style transfer approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 322–329. ACM.
- [137] Šimko, J. and Bieliková, M. (2014). State-of-the-art: Semantics acquisition and crowdsourcing. In *Semantic Acquisition Games*, pages 9–33.
- [138] Sohn, K., Shang, W., and Lee, H. (2014). Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149.
- [139] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.
- [140] Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge.
- [141] Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.
- [142] Suveeranont, R. and Igarashi, T. (2010). Example-based automatic font generation. In *International Symposium on Smart Graphics*, pages 127–138. Springer.
- [143] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- [144] Tenenbaum, J. B. and Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283.
- [145] Tewell, J., Bird, J., and Buchanan, G. R. (2017). The heat is on: A temperature display for conveying affective feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1756–1767. ACM.
- [146] Truillet, P., Oriola, B., Nespoulous, J.-L., and Vigoroux, N. (2000). Effect of sound fonts in an aural presentation. In *6th ERCIM Workshop, UI4ALL*, pages 135–144. Citeseer.
- [147] Tsonos, D. and Kouroupetroglou, G. (2016). Prosodic mapping of text font based on the dimensional theory of emotions: a case study on style and size. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1):1–16.
- [148] Tullis, T. S., Boynton, J. L., and Hersh, H. (1995). Readability of fonts in the windows environment. In *Conference companion on Human factors in computing systems*, pages 127–128. ACM.
- [149] Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 474–477. IEEE.
- [150] Walther, J. B., Loh, T., and Granka, L. (2005). Let me count the ways the interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of language and social psychology*, 24(1):36–65.
- [151] Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., and Li, J. (2014). Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166. ACM.

- [152] Wang, S. and Wang, X. (2005a). Emotion semantics image retrieval: an brief overview. In *Affective Computing and Intelligent Interaction*, pages 490–497.
- [153] Wang, S. and Wang, X. (2005b). Emotion semantics image retrieval: An brief overview. In *International Conference on Affective Computing and Intelligent Interaction*, pages 490–497. Springer.
- [154] Wang, Z., Yang, J., Jin, H., Shechtman, E., Agarwala, A., Brandt, J., and Huang, T. S. (2015a). Deepfont: Identify your font from an image. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 451–459. ACM.
- [155] Wang, Z., Yang, J., Jin, H., Shechtman, E., Agarwala, A., Brandt, J., and Huang, T. S. (2015b). Deepfont: Identify your font from an image. In *Multimedia*, pages 451–459. ACM.
- [156] Wierzbicka, A. (1986). Human emotions: universal or culture-specific? *American anthropologist*, 88(3):584–594.
- [157] Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250.
- [158] Williams, C. E. and Stevens, K. N. (1981). Vocal correlates of emotional states. *Speech evaluation in psychiatry*, pages 221–240.
- [159] Xie, J., Hertzmann, A., Li, W., and Winnemöller, H. (2014). Portraitsketch: Face sketching assistance for novices. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 407–417. ACM.
- [160] Xu, S., Lau, F. C., Cheung, W. K., and Pan, Y. (2005). Automatic generation of artistic chinese calligraphy. *IEEE Intelligent Systems*, 20(3):32–39.
- [161] Xu, Y. (2013). Prosodypro—a tool for large-scale systematic prosody analysis. Laboratoire Parole et Langage, France.
- [162] Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.
- [163] Yan, R. (2016). i, poet: Automatic poetry composition through recurrent neural networks with iterative polishing schema. In *IJCAI*, pages 2238–2244.
- [164] Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM.
- [165] Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. (2017a). Midinet: A convolutional generative adversarial network for symbolic-domain music generation using 1d and 2d conditions. *arXiv preprint arXiv:1703.10847*.
- [166] Yang, S., Liu, J., Lian, Z., and Guo, Z. (2017b). Awesome typography: Statistics-based text effects transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7464–7473.
- [167] Yao, B. (2011). *Mental simulations in comprehension of direct versus indirect speech quotations*. PhD thesis, University of Glasgow.
- [168] Yao, B., Belin, P., and Scheepers, C. (2011). Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *Journal of Cognitive Neuroscience*, 23(10):3146–3152.

- [169] Yeo, K. P. and Nanayakkara, S. (2013). Speechplay: composing and sharing expressive speech through visually augmented text. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, pages 565–568. ACM.
- [170] Zamani, H. and Croft, W. B. (2017). Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514. ACM.
- [171] Zhang, J., Wang, S., and Huang, Q. (2017). Location-based parallel tag completion for geo-tagged social image retrieval. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):38.
- [172] Zhang, Y., Zhang, Y., and Cai, W. (2018). Separating style and content for generalized style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1.
- [173] Zhao, S., Yao, H., Yang, Y., and Zhang, Y. (2014). Affective image retrieval via multi-graph learning. In *ACM MM*, pages 1025–1028.
- [174] Zramdini, A. and Ingold, R. (1998). Optical font recognition using typographical features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):877–882.
- [175] Zramdini, A. W. and Ingold, R. (1993). Optical font recognition from projection profiles. *Electronic Publishing*, 6(3):249–260.