Acta Technica *62* (2017), No. 7A, 335–346

A visual image design method based on three dimensional convolutional neural network in digital media environment

Jingzhen Cao[1]

Abstract. As vision has great impact on image design, image design should be varied with digital media context of product in the process of product design. In order to better realize visual image design, a kind of visual image design based on three-dimensional convolutional neutral network under the context of digital media was proposed in the thesis. Three-dimensional convolutional neutral network was introduced. In terms of all visual behaviors required for training, long-term visual behavior information was extracted as its senior visual behavior features. As time of the dynamic information is long, it should include more abundant information that cube of design factor for input image of convolutional neutral network. Then, convolutional neutral network was forced to learn a feature vector which really approaches the feature. A series of auxiliary input nodes were connected through the last implicit layer of convolutional neutral network so as to make extracted feature better approach the feature vector of visual behavior for calculated high layer in the training process. Finally, effectiveness of mentioned scheme is testified through case analysis.

Key words. Convolutional neural network, Three dimension, Visual image, Digital media context, Product design

.

1. Introduction

Visual communication has become the most important communication method in information society. Occurrence of new media directly influences visual communication in concept and method, thus visual form and feature, communication media and content, way to convey image, creation method of producer, and feeling and experience of consumer have changed greatly and innovation will be made with development of new media technology. Marshall McLuhan, a Media Theorist, considers media as "extension of human beings". If communication of spoken language is auditory extension of human beings; print media is visual extension of human

[1] Wuxi Institute of Arts and Technology, Jiangsu Yixing, 214206

beings; electronic media is central nervous extension of human beings; then, current new digital media is a media taking Internet as center. Emerging network media expands communication method of traditional media and extends comprehensive perception of human beings in real-time interaction, nonlinearity, hyperlink, and other several aspects. McLuhan makes people deeply understand great power of mass media in information society.

New media in a certain sense is a new language, a code to accumulate experience, and a visual machine. New media forms our new understanding to the actual world. Survival way of human beings determines basic position and important function of sense and sense experience for human beings. In western high visualized environment, meaning of "see" usually can be exchanged with the word "know". For example, people usually say "I see" rather than "I know". Hegel pointed that only vision and hearing among all senses of human beings are cognitive. Obviously, hearing plays basic and dominant role in all sensory organs. Information amount obtained through vision accounts 70% of obtained total amount; hearing accounts for 20%; obtained information amount of other senses only accounts for 10%. In addition, touch, hearing, smell, taste, and other senses with purpose of human beings have to be directed by vision. All visual behaviors of human beings start from cognition. Human beings have to experience three cognitive stages in the process of accepting complete visual experience: 1. Seeing. Feeling of audience to image starts from seeing. It is visual shock emphasized by us to catch the eye. 2. Watching. Attention of audience will turn from form of image to connotation of image. The more information amount and emotional factors cover in the image, the long the concern and reading time will be, thus the communication effect will be better. 3. Cognition. Cognitive stage is the highest stage of image reading. After obtaining image information, audience will conduct in-depth understanding to the image through association according to his cultural cultivation and life experience. Seeing, watching, cognition, and others of any visual information require brain to participate in visual behaviors, which is the process to achieve purposes through "thinking". What "seeing" emphases is interest point in a short time. Various compositions of design factor, form, and image for visual image can trigger excitement of vision, which can draw audiences' attention to "see". Interest of "seeing" for vision is the first factor required to be concerned at the time of all information communications. However, it is insufficient. "Watching" of audiences is necessary. Watching is a kind of selected visual behavior, which can cause understanding, memory, and thinking stage process of human beings. The result of watching is that human beings can incorporate what we see into what we can touch. When there are lots of opinions and selections about cause of "seeing" and when it is easy to trigger seeing, "watching" is no doubt a kind of rare and luxury visual behavior. Although we spend lots of time watching lots of things on Internet with complicated and numerous information in arbitrarily changed visual experience with clicking, nothing may be seen, because we do not "watch" any of them.

A kind of design method of visual image based on three-dimensional convolutional neutral network under the context of digital media is proposed in the thesis. Three-dimensional convolutional neutral network under the context of digital media is proposed in the thesis. Three-dimensional convolutional neutral network is introduced. In terms of all visual behaviors required for training, long-term visual behavior information is extracted as its senior visual behavior features so as to make extracted feature better approach the feature vector of visual behavior for calculated high layer in the training process.

2. Description of scheme

The concept of three-dimensional virtuality of context is shown to users through lingo displayed on a window (on screen of computer). User interface is used in the strategy to simulate the interaction mode of targets in reality. In addition, the position of background and concept can be displayed on the screen through (position, rotation, and scaling) camera at the interface according to the angle of head simulation of user. The camera on the screen can be used to obtain face position of user and Haar feature of cascade classifier. The algorithm can be realized in computer vision library of open source.

The first step is to obtain head position (x, y) of user so as to realize position of virtual camera in three-dimensional software. In addition, head should be placed on the same place, which is shown in Fig. 1. Distance z between virtual camera and virtual concept is added value for distance z_r between virtual camera and screen and distance z_c between screen and three-dimensional virtual concept. The distance is calculated on the basis of head diameter. As a result, a media $H = 156.214$mm is set so as to conduct measurement research for people at the age of 26 or 27.

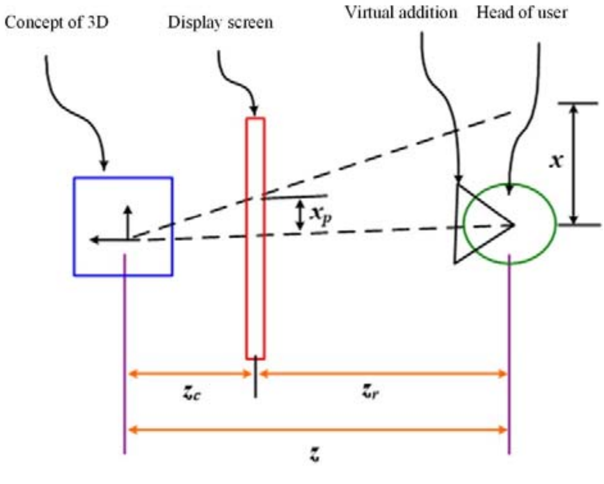


Fig. 1. Conceptual description of the system

Visual library includes a training instrument and a detector. In the case, classifier in software application which has been trained should be used. Classifier should be applied to interested region and should be used in mobile window so as to conduct facial search to input image after training of classifier. When human face is found, the algorithm will return to the position where human face is detected. Where (x_h, y_h) indicates the highest point of rectangle surrounding face; (ω, h) indicates width and height of rectangle.

Projection position (x_p, y_p) of head for user on the screen can be calculated as follows:

$$(x_p, y_p) = \left(\frac{x_h + \omega}{2}, \frac{y_h + h}{2}\right) . \tag{1}$$

Return value of all frames can be expressed as:

$$P\left(t\right) = \left(x_p, y_p, h\right) . \tag{2}$$

By taking advantages of similar triangle theorem, virtual concept position on the screen can be expressed as:

$$(x, y) = \left(\frac{x_p\left(z_c + z_r\right)}{z_c}, \frac{y_p\left(z_c + z_r\right)}{z_c}\right) . \tag{3}$$

Required variable is the corresponding position of head for user to camera, which is shown as follows with similar triangle theorem:

$$z_r = \frac{H \cdot f - h \cdot f}{h} . \tag{4}$$

Where f indicates focal length of IP camera. In addition, vector of virtual camera always point at center of three-dimensional virtual concept, thus shooting angle of virtual camera is calculated according to position of virtual concept and virtual camera.

3. Architecture of three-dimensional convolutional neural network

3.1. Three-dimensional convolution

A simple method to use convolutional neural network in video is to identify design elements of all images with convolutional neural network, but dynamic information between design elements of continuous images is not considered in the method. A kind of three-dimensional convention method is proposed in the thesis for effective and comprehensive dynamic information. Features which have distinctiveness in time dimension and space dimension are captured through conducting three-dimensional convolution in convolutional layer of convolutional neural network s. Three-dimensional convolution is used to organize a cube through piling up design elements of several continuous images and to use three-dimensional convolution kernel in the cube.

In the structure, all feature maps in the convolutional lay will be connected with design elements of several nearby continuous images in last layer so as to capture dynamic information. For example, in left upper image, value of certain position for a convolutional map is obtained through local receptive field of the same position for design elements of three continuous images in last convolutional layer. What needs to be concerned is that three-dimensional convolution kernel can only be used to extract a kind of feature in the cube, because weight of convolution kernel in the whole cube is identical, which means that the cube shares weight and has the same convolution kernel (the connection line of design element for the same visual image indicates identical weight). Several kinds of convolution kernel can be used to extract several features. In terms of convolutional neural network s, there is a common design ruler: No. of feature map for latter layer (which is close to input layer) should be increased, thus more kinds of features can be generated from lower feature maps combination.

3.2. Description of model

Architecture of three-dimensional convolutional neural network includes a hardwired layer, three convolutional layers, 2 sub-sampling layers, and a full connection layer. Convolutional cube of all three-dimensional convolution kernels are design elements of 7 continuous images. Size of design elements patch for all images is 60x40.

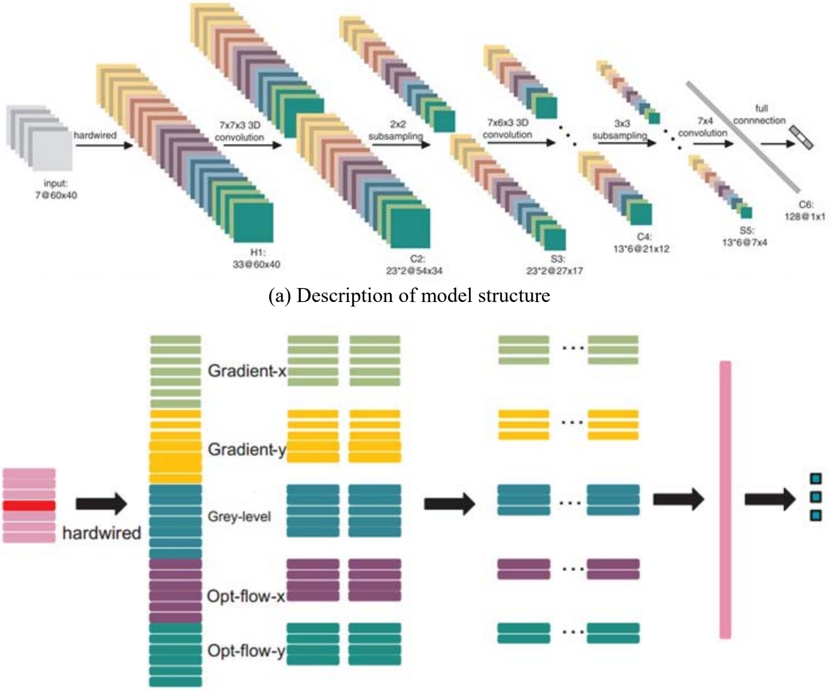
A fixed hardwired kernel in the first layer is used to process design elements of original image so as to generate information of several channels. Then, several channels should be separately processed. Finally, information of all channels should be combined so as to obtain final feature description. In fact, our transcendental knowledge to feature is encoded in the hardwired layer, which has better performance than that of random initialization.

Information of five channels should be extracted from design elements of all images, which separately is gray scale, gradient in direction x and y, and optical flow of direction x and y. The first three can be calculated through design elements of all images. However, optical flow field in horizontal and vertical direction can be determined through design elements of two continuous images, thus it has 7x3 + (7-1) x 2=33 feature maps.

Then, a 7x7x3 three-dimensional convolution kernel (7x7 is in the space; 3 is time dimension) will be used to separately conduct convolution in five channels. In order to increase No. of feature maps (which actually is to extract different features), two different convolution kernels are used at the same position, thus all groups in two groups of feature maps on C2 layer will include 23 feature maps. 23 is (7-3+1) x3 + (6-3+1) x2; the front one indicates design elements of seven continuous images; three channels of gray level, gradient in direction x and y will separately have 7 image design elements, while optical flow fields in horizontal and vertical direction only have 6 image design elements. 54x34 is (60-7+1) x (40-7+1).

In terms of max pooling in layer S3 of next sub-sampling layer, window 2x2 is used in feature maps of layer C2 to conduct sub-sampling so as to obtain feature maps with the same No. and lower spatial resolution. 27x17=(52/2)*(34/2) can be obtained after sub-sampling.

Three-dimensional convolution kernel 7x6x3 is separately used by C4 in 5 channels. In order to increase No. of feature maps, 3 different convolution kernels are used at all positions so as to obtain 6 groups of different feature maps with each group of 13 feature maps. 13 is ((7-3+1)-3+1)x3+((6-3+1)-3+1)x2; the front one indicates design elements of seven continuous images; three channels of gray level, gradient in direction x and y will separately have 7 image design elements, while optical flow fields in horizontal and vertical direction only have 6 image design elements. 21x12 is (27-7+1)x(17-6+1).



(a) Description of model structure

(b) Information processing of five channels

Fig. 2. Architecture of three-dimensional convolutional neural network

3.3. Model regularization

In terms of all visual behaviors required for training, long-term visual behavior information is extracted as its senior visual behavior features. As time of the dynamic information is long, it includes more abundant information that cube of design factor for input image of convolutional neutral network. Then, convolutional neutral network will be forced to learn a feature vector which really approaches the feature. A series of auxiliary input nodes can be connected through the last implicit layer of convolutional neutral network so as to make extracted feature better approach the feature vector of visual behavior for calculated high layer in the training process.

In the experiment, dense sift descriptor is calculated in original gray image; then, bag-of-words feature will be constructed as auxiliary feature through combination of these sift descriptors and motion edge history image (MEHI).

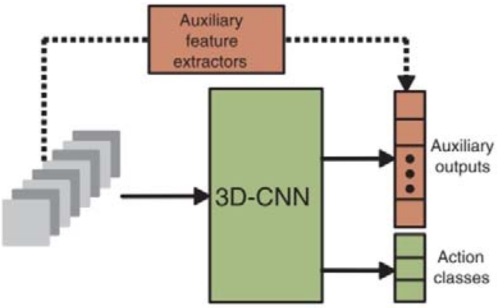


Fig. 3. Model regularization

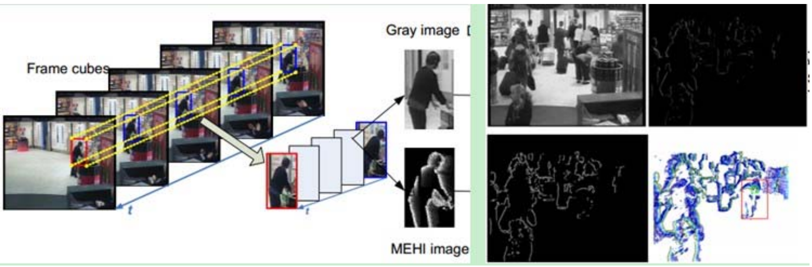


Fig. 4. Calculation of auxiliary features

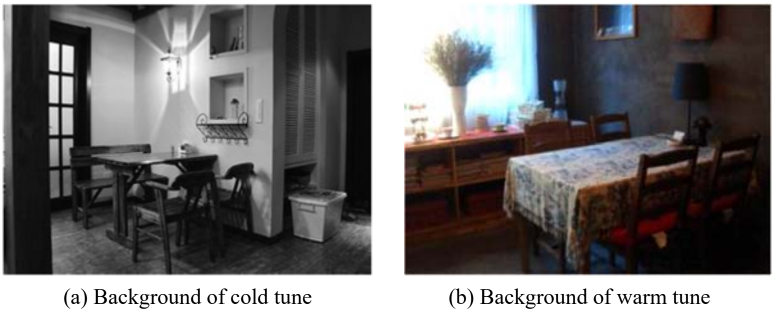
As appearance information is kept in gray image and only form and dynamic mode are concerned in MEHI, thus two pieces of complementary information is extracted as local feature bag of design elements for two continuous images. Calculation of MEHI is shown in right upper Fig. Differences between two image design elements are firstly and simply calculated so as to make observation image be clearer and more concise. Finally, these historical images should multiply by a forgetting factor and their results should be added so as to obtain overall MEHI.

4. Case analysis

4.1. Experiment setting

In order to evaluate function of NUI, two different cases in a design are researched. In terms of all conditions, 5 kinds of design elements for visual images should be set for lamp; design element combination for visual image which is suitable for user-defined specific background should be selected as well. The lamp is composed of part combination for 2 groups of different design elements for visual images.

In addition, there is a different target user with the intention to select the opposite style in all cases. One is male target user (in condition 1), the other is female target user (in condition 2). Both two target users are 22-year-old undergraduates. In terns of background for male target user, it is subject to cold tune; the form is closer to linearity under the element background. On the contrary, under the background of female target user, warm design elements for visual images are used as background, which is shown in Fig. 5 in detail.


(a) Background of cold tune

(b) Background of warm tune

Fig. 5. Two different design environments

Per target user shall participate in four kinds of selection process for design element of visual image, and it is shown in Table 1. Different combinations shall be formulated, and they shall be arranged from selection to select the favorite combination.

Table 1. Experimental color symbols

Circumstance 1 (male)		Circumstance 2 (female)	
Code	RGB	Code	RGB
n	0,0,0	n	0,0,0
b	255,255,255	b	255,255,255
a	0,71,255	r	197,0,11
v	92,133,38	c	76,25,0

In addition, description of target user was printed and handed over to per case for design. It included users' personal data, favorite visual image design element, film, place, hobby, brand and background where the light in the picture is in. Three kinds of users were considered in the experiment:

(1) Target user, who was the purchase and user of product without three-dimensional modeling experience. (2) Design Team, which is composed of Product Design Engineer. Every designer was required to confirm design element combination for five visual images to better describe specific users. (3) Expert Team, which was composed of professors in different technical fields.

4.2. Result and discussion

In order to analyze interaction for these two interfaces, 18 designers recorded and interacted with software. At the same time, there was no time limit, and emotional state or design skill for per designer was out of control. For interaction with NUI, 100% user can fully interact with interface with ease. But for typical interface interaction, 55% users think that this interface is more dispersed and difficult to be operated, which is shown in Fig.6.

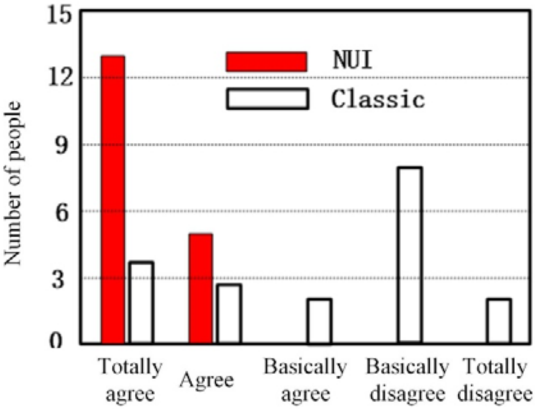


Fig. 6. Difficulty comparison for interface interaction

In order to make selection for visual image design element in decision-making activity easier, 94% designers think that NUI has advantage. But generally speaking, no matter whether it is typical interface or NUI, they both can simplify design element selection for visual image. Specific condition is shown in Fig.7.

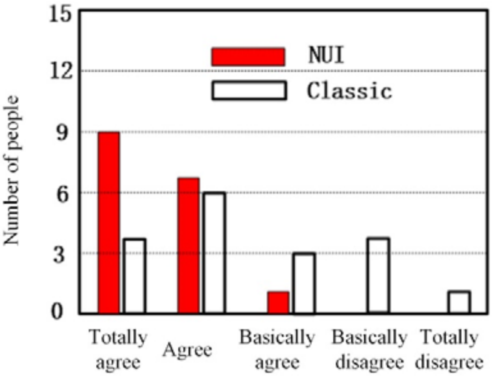


Fig. 7. Difficulty comparison for design element selection of visual image

It can be known from experimental result in Fig. 6-7 that NUI adopted in the Thesis for selection scheme for visual image design element is more recognized by evaluation team compared with traditional typical selection scheme for visual image design element for evaluation team made up of three different kinds of users, which represented effectiveness for proposed method.

5. Conclusion

One kind of visual image design method based on three-dimensional convolutional neural network under the context of digital media was proposed in the Thesis. During the process of product design, it should be different according to different digital contexts where products were in, and three-dimensional convolutional neural network was introduced. Visual behavior information for long time was extracted to be as its advanced visual behavior feature. Then through the last hidden layer for convolutional neural network, a series of auxiliary output node was connected to make better approximation of extracted feature to feature vector of advanced visual behavior calculated. Effectiveness of the proposed scheme was verified in case analysis result.

Acknowledgement

Jiangsu University Philosophy and social science research project (guidance) – Research on the inheritance and dissemination of traditional culture of purple sand under digital media environment (2015SJD390).

References

[1] Y. Y. Zhang, J. W. Chan, A. Moretti, and K. E. Uhrich: *Designing Polymers with Sugar-based Advantages for Bioactive Delivery Applications*, Journal of Controlled Release, *219* (2015), 355–368.

[2] Y. Y. Zhang, Q. Li, W. J. Welsh, P. V. Moghe, and K. E. Uhrich: *Micellar and Structural Stability of Nanoscale Amphiphilic Polymers: Implications for Anti-atherosclerotic Bioactivity*, Biomaterials, *84* (2016), 230–240.

[3] E. W. Chan, Y. Y. Zhang, and K. E. Uhrich: *Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation*, Bioconjugate Chemistry, *26* (2015), No. 7, 1359–1369.

[4] Y. Y. Zhang, E. Mintzer, and K. E. Uhrich:*Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes*, Journal of Colloid and Interface Science, *482* (2016), 19–26.

[5] Y. Y. Zhang, A. Algburi, N. Wang, V. Kholodovych, D. O. Oh, M. Chikindas, and K. E. Uhrich: *Self-assembled Cationic Amphiphiles as Antimicrobial Peptides Mimics: Role of Hydrophobicity, Linkage Type, and Assembly State*, Nanomedicine: Nanotechnology, Biology and Medicine, *13* (2017), No. 2, 343–352.

[6] J. Jiang, P. Trundle, J. Ren: *Medical image analysis with artificial neural networks*[J]. Computerized Medical Imaging and Graphics, *34* (2010), No. 8, 617–631.

[7] S. Ji, W. Xu, M. Yang, et al.: *3D convolutional neural networks for human action recognition*[J]. IEEE transactions on pattern analysis and machine intelligence, *35* (2013), No. 1, 221–231.

[8] H. Yan: *Image analysis for digital media applications*[J]. IEEE Computer Graphics and Applications, *21* (2001), No. 1, 18–26.

[9] H. Yoshida, W. Zhang, R. M. Nishikawa, et al.: *Method for determining an optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms: U.S.* Patent 6,075,878[P], (2000). 2000-6-13.

[10] J. C. McCall, M. M. Trivedi: *Video-based lane estimation and tracking for driver*

assistance: survey, system, and evaluation[J]. IEEE transactions on intelligent transportation systems, *7* (2006), No. 1, 20–37.

[11] R. Nekovei, Y. Sun: *Back-propagation network and its configuration for blood vessel detection in angiograms*[J]. IEEE Transactions on Neural Networks, *6* (1995), No. 1, 64–72.

[12] H. R. Roth, L. Lu, J. Liu, et al.: *Improving computer-aided detection using convolutional neural networks and random view aggregation*[J]. IEEE transactions on medical imaging, *35* (2016), No. 5, 1170–1181

[13] C. Segalin, D. S. Cheng, M. Cristani: *Social profiling through image understanding: Personality inference using convolutional neural networks*[J]. Computer Vision and Image Understanding (2016).

[14] M. H. Yang, D. J. Kriegman, N. Ahuja: *Detecting faces in images: A survey*[J]. IEEE Transactions on pattern analysis and machine intelligence, *24* (2002), No. 1, 34–58.

Received May 7, 2017