附件 6-4 中心近三年发表的科研论文清单

序号	作者	论文题名	刊物名称	发表卷期	刊物级别	发表时间
1	Changhong Chen,Hehe Dou, Zongliang Gan	Collective Activity Recognition by Attribute-based Spatio-temporal Descriptor(SCI: 000364422400019)(EI: 20154101345539)	IEICE Trans. on Information and Systems	2015年01期	SCI检索	2015.01
2	Dandan Si, Yuanyuan Hu, Zongliang Gan, Ziguan Cui, Feng Liu	Edge Directed Single Image Super Resolution Through the Learning Based Gradient Regression Estimation(EI: 20154201380942)	The 8th International Conference on Image and Graphics	2015年08期	EI检索	2015.01
3	Guijin Tang, Xiaohua Liu, Changhong Chen, Lei Wang, Ziguan Cui1, Zongliang Gan, Feng Liu	Active Tracking Using Color Silhouettes for Indoor Surveillance(EI: 20160902035330)	The International Conference on Wireless Communications and Signal Processing	2015年01期	EI检索	2015.01
4	Ziguan Cui, Zongliang Gan, Guijin Tang, Feng Liu, Xiuchang Zhu	Image Signature Based Mean Square Error for Image Quality Assessment(SCI:000362918000015)(EI: 20154001337480)	Chinese Journal of Electronics	2015年01期	SCI检索	2015.01
5	单美贤	CSCL协作问题解决过程中的学习支持工具研究综述	电化教育研究	2015年01期	CSSCI	2015.01
6	刘峰,施阳,于宗良，秦雷，陈昌红	基于BJND和JNDD的立体视频深度感知增强技术综述	南京邮电大学学报（自然科学版）	2015年01期	北大核心	2015.01
7	Danfeng Zhao, Yuanyuan Hu, Zongliang Gan.Changhong Chen,and Feng Liu	A Novel Improved Binarized Normed Gradients Based Objectness Measure Through the Multi-feature Learning(EI: 20154201380461)	The 8th International Conference on Image and Graphics	2015年08期	EI检索	2015.04
8	Feng Liu, Ruoxuan Yin, Zongliang Gan, Changhong Chen, Guijin Tang	Robust Face Hallucination via Similarity Selection and Representation(EI: 20154201381933)	The 8th International Conference on Image and Graphics	2015年08期	EI检索	2015.04
9	张刚要	基于TAM的高校网络课程接受度影响因素研究	南京邮电大学学报	2015年05期	北大核心	2015.05

10	章宁	我国成人高等教育的困境与战略转型	成人教育	2015年07期	北大核心	2015.07
11	单美贤	CSCL合作问题解决过程中情感支持框架探析	开放教育研究	2015年04期	CSSCI	2015.08
12	单美贤	美国高等院校有效教学的调查与分析:以佛罗里达大学为例	高教探索	2015年09期	CSSCI	2015.09
13	崔子冠,干宗良,唐贵进，刘峰,朱秀昌	联合梯度强度与方向信息的图像质量评价	仪器仪表学报	2015年12期	EI检索	2015.12
14	Baibai Xu, Changhong Chen, Zongliang Gan	Aurora sequences classification and aurora events detection based on hidden conditional random fields(EI: 20164703028127)	Chinese Conference on Pattern Recognition	2016年01期	EI检索	2016.01
15	Peiqing Bai, Ziguan Cui, Zongliang Gan,Guijin Tang,Feng Liu	A Novel Saliency Detection Model Based on Curvelet Transform(EI: 20165203172882)	The International Conference on Wireless Communications and Signal Processing	2016年01期	EI检索	2016.01
16	陈媛媛	“AVAILABLE SEAT?”Case analysis of interactive device and interactive design project	16th IEEE International Conference on Ubiquitous Wireless Broadband，ICUWB 2016(EI Compendex)	2016,Volume 89	EI检索	2016.01
17	刘峰	Online Video Synopsis Methid through Simple Tube Projection Strategy	WCSP 2016	2016年01期	EI检索	2016.01
18	刘峰	Vehicle Type Classification via Adaptive Feature Clustering for Traffic Surveillance Video	WCSP 2016	2016年01期	EI检索	2016.01
19	刘峰	Enhancement of Color Image Based on Tone-Preserving	2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)	WCSP 2016	EI检索	2016.01
20	王克祥	反思无名政治受难者纪念碑竞赛	装饰，ISSN：0412-3662	2016年10期	北大核心	2016.01
21	Guo-gang Wang, Gui-jin Tang , Zong-liang Gan , Zi-guan Cui , Xiu-chang Zhu	Basic Problems and Solution methods for Two-Dimensional Continuous 3×3 Order Hidden Markov Model(SCI:000378450200050)(EI: 20154001337480)	Chaos Solitons & Fractals		SCI	2016.03

22	花景培,陈昌红,干宗良，刘 峰	基于运动和外形度量的多目标行人跟踪	南京邮电大学学报（自然科学版）	2016年01期	北大核心	2016.03
23	刘思江	One-click scanning of large-size documents using mobile phone camera	会议论文，预计EI收录		EI检索	2016.04
24	卢锋	教育技术学期刊国际影响力提升策略研究	科技与出版，ISSN：1005-0590	2016年第6期	CSSCI	2016.06
25	林巧民	A method of cleaning RFID data streams based on Naive Bayes classifier	IJAHUC,ISSN:1743-8225，WOS:000376173100003，INDERSCIENCE ENTERPRISES LTD，SCI收录	2016年21期	A类	2016.07
26	卢锋	媒介素养教育的发展动因研究	现代远距离教育，ISSN：1001-8700	2016年第4期	CSSCI	2016.08
27	卢锋	文化向度的国际媒介素养教育考察	现代传播，ISSN：1007-8770	2016年第8期	CSSCI	2016.08
28	陈媛媛	“AVAILABLE SEAT?” Case analysis of interactive device and interactive design project	ICUWB2016		EI检索	2016.1
29	Jin Jin, Feng Liu, Zongliang Gan, Ziguan Cui	Online Video Synopsis Methid through Simple Tube Projection Strategy(EI:20165203172945)	Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on IEEE	2016年11期	EI检索	2016.11
30	Shu Wang, Feng Liu, Zongliang Gan, Ziguan Cui	Vehicle Type Classification via Adaptive Feature Clustering for Traffic Surveillance Video(EI:20165203172571)	Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on. IEEE	2016年11期	EI检索	2016.11
31	Ya’nan Yang, Xiaofan Wang, Feng Liu, Zongliang Gan	Enhancement of Color Image Based on Tone-Preserving(EI:20165203172521)	Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on IEEE	2016年11期	EI检索	2016.11
32	王克祥	“图说画映”——苏州桃花坞木刻年画叙事性研究	艺术百家，ISSN：1003-9104	2016年06期	北大核心	2016.11
33	陈媛媛	公共空间新媒体艺术观念与表现研究	美术与设计（南京艺术学院学报）	2016年第6期	CSSCI、北大核心来源期刊	2016.12
34	余武	基于自适应学习系统的个性化图书推荐研究	《山西青年》	2017年01期	省刊	2017.01
35	张刚要	教育回归生活世界：技术具身性的启示	当代教育科学	2017年01期	北大核心	2017.01

36	李峻	“民国工匠”培养的个案分析与当代反思——以“国立第一职业学校”为中心	职业技术教育	2017年02期	北大核心	2017.02
37	单美贤	网络文化安全治理的国际经验探析	南京邮电大学学报（社会科学版）	2017年01期	北大核心	2017.03
38	李峻	我国农民工继续教育政策存在的问题与优化路径——基于“求学圆梦行动”的文本分析	职教论坛	2017年03期	北大核心	2017.03
39	余武	Moos,翻转课堂教学模式在高校影视类课程中的创新探索	兴义民族师范学院学报	2017年01期	省刊	2017.03
40	张刚要	教学媒体：由技术工具论、工具实在论到具身理论的范式转换	中国电化教育	2017年04期	CSSCI	2017.04
41	刘永贵	走向2030:中国高等教育现代化建设之路	中国高教研究	2017年05期	CSSCI	2017.05
42	姜玻	Gaze inspired subtitle position evaluation for MOOCs videos	IWPR 2017(SPIE Proceedings Volume 10443)	2017,Volume 10443	EI会议	2017.06
43	姜玻	Interactive QR code beautification with full background image embedding	IWPR 2017(SPIE Proceedings Volume 10443)	2017,Volume 10443	EI会议	2017.06
44	刘思江	Deep learning application: rubbish classification with aid of an android device	Proc. SPIE 10443,	2017,Volume 10443	EI检索	2017.06
45	章宁	民国初期平民教育对新型职业农民培养的启示	成人教育	2017年06期	北大核心	2017.06
46	霍智勇	基于稳态匹配概率的光照鲁棒立体匹配算法的研究	南京邮电大学学报(自然科学版)	2017年08期	省刊	2017.08
47	唐湘宁	校企合作:成人高等教育办学机制转型的现实选择	成人教育	2017年09期	北大核心	2017.09
48	徐水晶	教育作为阶层代际传递的中介作用研究	社会科学	2017年09期	CSSCI	2017.09
49	李峻	大学跨学科学术组织的成长逻辑与创新策略	江苏高教	2017年10期	cssci	2017.1
50	霍智勇	基于多种边缘暗示和尺度修正的RGB-D图像层次分割	南京邮电大学学报(自然科学版)	2017年11期	省刊	2017.11
51	李峻	路易·艾黎的创造性职业教育思想与实践	成人教育	2017年12期	北大核心	2017.11
52	唐湘宁	行业性院校的治理困境与创新逻辑	江苏高教	2017年11期	CSSCI	2017.11
53	杨祥民	笔墨引领时代——新时期中国画艺术发展的新命题	艺术百家	2015年第4期	CSSCI、北大核心来源期刊	2015/7/15
54	杨祥民	18世纪末英国家具设计“谢尔顿风格”成因及其特点	装饰	2015年第8期	CSSCI、北大核心来源期刊	2015/8/15

55	袁潇	美国广告学教育应对数字时代的启示	传媒观察(曾用刊名：新闻通讯)	2015 年第 9 期	北大核心来源期刊	2015/9/10
56	季静	游戏与奇观：《最强大脑》节目元素的文化分析	传媒观察(曾用刊名：新闻通讯)	2015 年 11 期	北大核心来源期刊	2015/11/10
57	袁潇	基于手机媒体使用的青少年亚文化族群研究	编辑之友	2016 年第 4 期	CSSCI、北大核心来源期刊	2016/4/5
58	袁潇	数字时代中议程设置理论的嬗变与革新	国际新闻界	2016 年第 4 期	CSSCI、北大核心来源期刊，《人大复印资料》全文转载	2016/4/23
59	季静	“极速前进”中的“极限挑战”——对 2015 年中国电视综艺节目的思考	中国电视	2016 年第 5 期	CSSCI、北大核心来源期刊	2016/5/15
60	杨祥民	略论人类早期扇子的形制	美术观察	2016 年第 5 期	CSSCI、北大核心来源期刊	2016/5/15
61	范建华	用美洗刷人心，用美浸润人生——陈之佛的艺术心路	美术观察	2016 年第 9 期	CSSCI、北大核心来源期刊	2016/9/15
62	范建华	在“写生”与“写死”中新生——五十年代“新国画运动”之演进	文艺理论与批评	2016 年第 5 期	CSSCI、北大核心来源期刊	2016/9/24
63	刘峰	Online Video Synopsis Methid through Simple Tube Projection Strategy	Wireless Communications & Signal Processing (WCSP), 2016 8th International Conference on. IEEE, 2016: 1-5		EI 检索	2016/11/1
64	吴婧婧	探究混合媒介视野下媒体广告的艺术创意及情感体验	学术论坛	2016 年第 12 期	CSSCI、北大核心来源期刊	2016/12/1
65	刘峰	基于主题相似度的视频分段	南京邮电大学学报(自然科学版)	2017 年第 6 期	北大核心	2017/1/5
66	袁潇	宫崎骏动画电影中的生态叙事观研究	传媒观察	2017 年第 1 期	北大核心、SCD 收录	2017/1/10
67	刘峰	基于部件时空信息的目标跟踪算法	计算机工程与设计	2017 年第 1 期	北大核心	2017/1/16
68	刘峰	基于压缩特征的尺度自适应目标跟踪算法	南京邮电大学学报(自然科学版)	2017 年第 1 期	北大核心	2017/3/2
69	袁潇	移动网络下集体行动的传播机制研究——基于 J 市出租车停运事件的个案考察	当代传播	2017 年第 2 期	CSSCI、北大核心收录	2017/3/15

70	吴婧婧	Analysis of the Impact of New Media Communication Based on Large Data Environment	REVISTA TECNICA DE LA FACULTAD DE INGENIERIA UNIVERSIDAD DEL ZULIA (Technical Journal of the Faculty of Engineering, TJFE, ISSN:0254-0770)		EI 收录	2017/4/1
71	刘峰	基于对数图像处理模型的低照度图像增强算法	南京邮电大学学报（自然科学版）	2017 年第 2 期	北大核心	2017/4/28
72	季静	电视剧网络口碑传播的特征及路径	编辑之友	2017 年第 6 期	CSSCI、北大核心收录	2017/6/5
73	季静	电视剧影响力评价标准刍议——从中国电视剧奖项说起	南京艺术学院学报（音乐与表演）	2017 年 3 期	北大核心收录	2017/8/15
74	刘峰	行人重识别研究综述	智能系统学报	2017 年第 6 期	北大核心	2017/11/9
75	单美贤	CPBL 教学法在本科教学中的实践分析：以交互设计课程为例	江苏高教	2017 年 03 期	CSSCI	2017.03
76	李峻	“一带一路”战略下我国边疆民族地区职业教育治理研究	社会科学家，2017,(01)20170125.	2017 年 01 期	北大核心	2017.01
77	李峻	“互联网+”背景下成人高等教育的发展困境与空间拓展	成人教育，2017，20170315	2017 年 03 期	北大核心	2017.03
78	刘峰	基于部件时空信息的目标跟踪算法	计算机工程与设计	2017 年 01 期	北大核心	2017.01
79	刘峰	基于 FPGA 的 1080P 低质视频实时增强系统	计算机技术与发展	2017 年 06 期	北大核心	2017.06

Building and Environment 121 (2017) 1–10

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

A pilot study of online non-invasive measuring technology based on video magnification to determine skin temperature

Xiaogang Cheng [a, b], Bin Yang [b, *], Thomas Olofsson [b], Guoqing Liu [c], Haibo Li [d]

[a] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] Department of Applied Physics and Electronics, Umeå University, Umeå, Sweden
[c] School of Physical and Mathematical Sciences, Nanjing Tech University, Nanjing, China
[d] School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 10 February 2017
Received in revised form
9 May 2017
Accepted 11 May 2017
Available online 12 May 2017

Keywords:
Thermal sensation
Skin temperature
Video magnification
Online non-invasive measurement

ABSTRACT

Much attention was paid on human centered design strategies for environmental control systems of indoor built environments. The goal is to achieve thermally comfortable, healthy and safe working or living environments in energy efficient manners. Normally building Heating, Ventilation and Air Conditioning (HVAC) systems have fixed operating settings, which can't satisfy human thermal comfort requirements under transient and non-uniform indoor thermal environments. Therefore, human thermal physiology signal such as skin temperature, which can reflect human body thermal sensation, has to be measured over time. Several trials have been performed by minimizing measuring sensors such as i-Button and mounting measuring sensors into wearable devices such as glasses. Infrared thermography technology has also been tried to achieve non-invasive measurements. However, it would be much more convenient and feasible if normal computer camera could record images, which could be used to obtain human thermal physiology signals. In this study, skin temperature of hand back, which has a high density of blood vessels and is normally not covered by clothing, was measured by i-button sensors. Images recorded by normal camera were amplified to analyzing skin temperature variation, which are impossible to see with naked eyes. The agreement between i-button sensor measuring results and image magnification results demonstrated the possibility of non-invasive measuring technology by image magnification. Partly personalized saturation-temperature model ($T = 96.5 \times S + b_i$) can be used to predict skin temperatures for young East Asia females.

1. Introduction

About 30% of total final energy has been consumed by commercial and residential buildings in US in recent years [1]. Higher percentage of final energy has been consumed in area with severe outdoor environments such as Singapore recently, which used 52% of electricity [2]. Heating, Ventilation and Air Conditioning (HVAC) systems account for largest portion of building energy consumption. Reduction of HVAC energy consumption is necessary while human thermal comfort stays a design goal, which is energy efficient thermal comfort. Based on the concept of human centered design and demand controlled conditioning, new strategies including displacement, stratum, task/ambient were developed. Non-uniform thermal environments were created by displacement ventilation [3], stratum ventilation [4], personalized ventilation [5,6], under floor air distribution [7], personal comfort system [8–10], etc.

Human thermal sensation, which can be reflected by thermal physiological parameters such as skin temperature, can be used as feedback signal to control above mentioned HVAC systems. Invasive measuring technology has been used for human skin temperature measurements, which is mainly for laboratory experiments [11]. Several trials have been performed by minimizing measuring sensors such as i-Button and mounting measuring sensors into wearable devices such as glasses [12]. However, it would be much more convenient and feasible if normal computer camera could record images, which could be used to obtain human thermal physiology signals. Human skin color changes slightly with vasodilation or vasoconstriction especially under local thermal stimuli such as using hand warmer. The variation, while invisible to the

* Corresponding author.
E-mail address: bin.yang@umu.se (B. Yang).

http://dx.doi.org/10.1016/j.buildenv.2017.05.021
0360-1323/

naked eyes, can be extracted by image magnification [13–18]. This study is focused on non-invasive measuring technology by image magnification for hand skin temperature measurements, which can be used for demand control of HVAC systems.

Human visual system has threshold values on temporal-spatial sensitivity. There are a lot of information, which can not be observed by human visual system and need magnification technique. Motion magnification technique was pointed out, which measured small motions by a robust analysis of feature point trajectories and segment pixels based on similarity of position, color and motion [13]. Very small motions were analyzed according to correlation over time. The technique, which acted like a microscope, can achieve magnified observations for tiny motions. One parameter cartoon animation filter was demonstrated to simultaneously add exaggeration, anticipation, follow-through, and squash and stretch to a wide variety of motions [14]. Based on video imaging and blind source separation, a method of non-contact, automated cardiac pulse measurements was introduced [15]. The results extracted from webcam based videos were compared with the results from finger blood volume pulse sensor. High accuracy was achieved. Red, green and blue (RGB) signals of skin color from human face were magnified and extracted the first time. Independent component analysis was used to remove noise and separate cardiac pulse, which achieved automated cardiac pulse measurements. One vital signs camera algorithm was presented, which magnified variation rate of skin color to achieve non-contact pulse and breathing rate measurements accurately [16].

Above mentioned methods followed Lagrangian perspective, which paid attention to motion trajectories of each pixel and were sensitive to tiny motions. However, accurate motion prediction and image segment technique made the algorithm complicated. The effect of different temporal sampling kernels was studied, which demonstrated extended overlapping kernels can mitigate aliasing artifacts [17]. Temporal processing was used to extract invisible signals [15]. Based on Eulerian perspective, Eulerian video magnification (EVM) algorithm was pointed out [18,19]. Eulerian spatial-temporal processing was used for monocular video sequences to magnify tiny variations, which can't be seen by naked eyes. EVM algorithm can magnify spatial channels and temporal channels respectively, which is suitable for magnifying color variations of image pixels under temporal channels. This was the first time to magnify video color and motion by Eulerian method. Phase-based video magnification approach was introduced to overcome the limit that only small magnification factors were supported at high spatial frequencies [20]. Layer-based video magnification approach was presented, which can amplify small motions within large motions [21]. An examined layer was temporally aligned and subtle variations were magnified. Matting was used to magnify only region of interest while maintaining integrity of nearby sites.

Because of sensitivity for color variations of image pixels under temporal channels, EVM will be used to correlate skin temperature and skin color saturation in this study. Linear relationship between skin temperature and skin color saturation is envisaged, which is the research hypothesis. It is the first time to use video magnification approach to determine skin temperature and thermal sensation, which can be used to control HVAC systems. Non-invasive measuring technology will be achieved.

2. Practical application

In private office and open plan office where personal computers were used for each staff, video of human naked skin can be recorded by computer camera and skin color saturation can be analyzed by video magnification technique (Fig. 1). After user identification, the cohort of each user was identified. Saturation-temperature (ST) models for different cohorts were recorded in computer database. As one pilot study, ST model for young East Asian females was analyzed. ST models for different cohorts will be analyzed, based on age, gender and race. The database of ST models for different cohorts will be developed. By inputting skin color saturation to corresponding ST model, user skin temperature can be calculated and used as feedback signal for HVAC system control. The concrete steps are as follows.

2.1. User identification

User identification can be performed by identifying some personal information such as personal computer user account, radio frequency identification, fingerprint, human face, etc.

2.2. Online study

For personal computer in each workstation, ST relationship for the fixed occupant can be sampled and analyzed, which is called initialized study. Skin saturation, obtained by video magnification and analysis, can be used to calculate skin temperature by ST

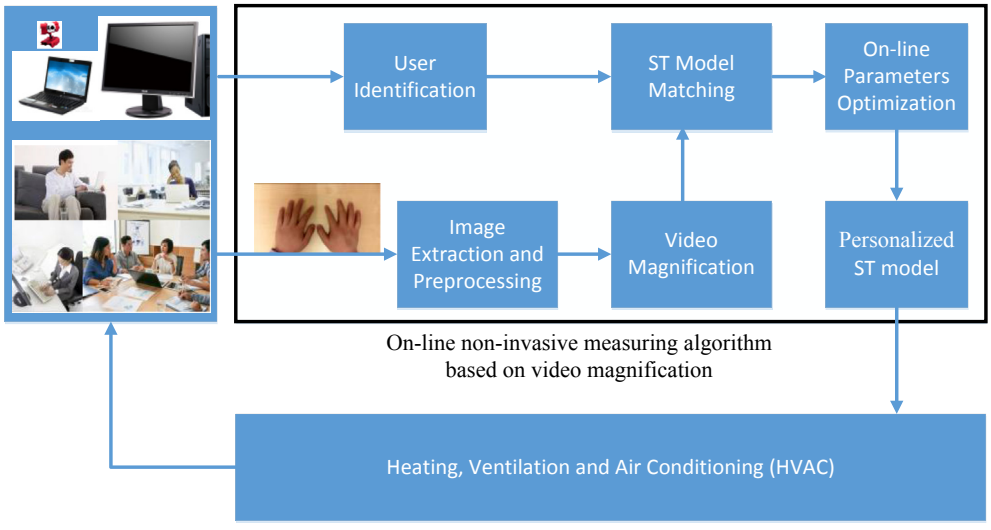


Fig. 1. Schematic of algorithm for practical application.

relationship for the fixed occupant. It is the most accurate method because ST relationship is the one for the studied person. The method can be used to control localized heating or cooling devices such as personal comfort system. For open plan office and conference room, ST relationships of each person should be integrated into one ST model for the cohort to control centralized HVAC system. For foreign visitors, ST model of their race can be found from database and used.

3. Research methods

3.1. Skin color saturation

Color space was defined artificially, which used three or four independent variables to describe colors. Commonly used color spaces included RGB (red, green, blue), CMYK (cyan, magenta, yellow, key (black)), YUV (Y: Luminance, UV: Chrominance), YIQ (Luminance, In-phase, Quadrature-phase) and HSV (Hue, Saturation, Value). Different color spaces had different advantages and limitations. Some of them can be converted by each other.

Space coordinate axis of HSV had inverted hexcone shape [22]. In counterclockwise direction, hue was the dimension with points on it normally called red, yellow, green, blue, magenta, etc. Value measured the departure of a hue from black, the color of zero energy. Saturation measured the departure of a hue from achromatic. The range of saturation is 0–1. The darker the color is, the higher the saturation value is.

The purpose of this study is to find non-invasive method for measuring human skin temperature, which can be used for HVAC system control. RGB signals of skin color from human face were magnified and extracted to analyze cardiac pulse [15]. However, RGB can not reflect saturation, which can be reflected by HSV. Pores expand and skin turns red when skin temperature increases. Skin color saturation may have close relationship with skin temperature. HSV was chosen for this study.

3.2. Thermal stimulus experiments

In indoor environments, people often experience thermal stimuli such as elevated air movement, asymmetric radiation, localized heating or cooling, etc. Under these thermal stimuli, local skin temperature may change which reflect changes of thermal sensation. As the first trial, strong stimulus to human hands was tried to find obvious results. 16 young East Asian females were chosen as human subjects. Young female subjects have relatively delicate skin, which has no skin folds and are sensitive to thermal stimulus. Their anthropometric data were shown in Table 1.

Experiments were performed in one environmental chamber with accurate temperature and humidity control. Dry-bulb temperature and relative humidity (RH) were continuously measured at one-minute sampling intervals by HOBO temperature/RH/light data loggers (Model U12−012, Onset, Bourne, Massachusetts, USA), with −20 to 70 °C measuring range, ±0.35 °C uncertainty for dry-bulb temperature; and 5%–95% measuring range, ±2.5% uncertainty for RH. Air speed was very low and radiant temperature was close to dry-bulb temperature. Experimental conditions were shown in Table 2.

Experiment for each subject lasted 60 min (Fig. 2). Subjects were asked to arrive to the test room 10 min before the experiment started. After taking off coats, clothing thermal resistance was kept at about 1 clo. When subjects just entered the room and were seated, they were asked to fill in the survey regarding anthropometric data. After 10 min thermal adaptation, their hands were immersed into 45 °C constant temperature water bath for 10 min. After that, hands were wiped dry and i-button sensors were pasted on hand backs immediately. Skin temperatures of hand backs were continuously measured for 50 min at one-minute sampling intervals by i-button (Model DS1921H, Maxim Integrated, San Jose, California, USA), with −30 to 70 °C measuring range, ±0.125 °C uncertainty for skin temperature. Videos were recorded for the 50 min by cell phone camera (Model G750-T00, 720P (1280 × 720), Huawei, Shenzhen, Guangdong, China). Corresponding images at the one-minute sampling intervals were extracted and analyzed. Matlab was used to do programming for video magnification. Subjects kept seated with about 1.1 met metabolic rates.

Table 1
Anthropometric data (mean ± standard deviation) of human subjects.

Gender	Sample size	Age (y)	Height (m)	Weight (kg)	BMI[a] (kg/m^2)
Female	16	23.9 ± 3.9	1.62 ± 0.05	52.2 ± 6.5	19.9 ± 2.2

[a] Body mass index = weight (kg)/[height (m)]^2.

3.3. ST model

$$V = \{f_1, f_2, \ldots, f_n\} \tag{1}$$

V − An video of hand skin sampling
f − Image frames in an video
n − Sampling time (s)

As mentioned above, saturation is the parameter that represents the color depth in HSV color space. In this study, a research hypothesis was proposed that linear relationship existed between skin color saturation and skin temperature.

$$T = aS + b \tag{2}$$

T − Skin temperature (°C)
S − Skin color saturation

where $T = \{T_1, T_2, \ldots, T_n\}$. Similarly,$a$, S, b are vectors. If the number of subjects is m, equation (2) can be expressed as follows.

$$\begin{bmatrix} T_{11} & T_{12} & \ldots & T_{1m} \\ T_{21} & T_{22} & \ldots & T_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ T_{n1} & T_{n2} & \ldots & T_{nm} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \times \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1m} \\ S_{21} & S_{22} & \ldots & S_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ S_{n1} & S_{n2} & \ldots & S_{nm} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \\ \vdots & \vdots & \ldots & \vdots \\ b_{n1} & b_{n2} & \ldots & b_{nm} \end{bmatrix} \tag{3}$$

The change in skin color saturation caused by skin temperature was extremely weak. To characterize saturation signal, the EVM technology [18] was introduced in this study.

$$\widehat{S}(x, y, t) = (1 + \alpha)S(x, y, t) + \varphi(x, y, t) \tag{4}$$

$\widehat{S}$ − Skin color saturation after amplification
S − Skin color saturation before amplification
x, y − Spatial coordinate

Table 2
Experimental conditions.

Target dry-bulb air temperature (°C)	Measured dry-bulb air temperature (°C)	Target relative humidity (%)	Measured relative humidity (%)
22	22.2 ± 0.2	40	36.9 ± 2.5

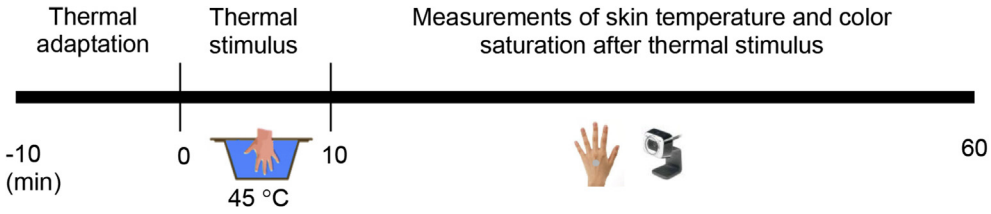


Fig. 2. Experimental procedure.

t − time
α − Amplification factor
φ − Gauss noise

$S\ (x, y, t)$ is image saturation at location (x, y) and time t. α is amplification factor. φ is noise, which is assumed to obey Gauss noise in the algorithm and will be removed by filtering. The details of the algorithm are as shown in Table 3 with detailed explanations followed. Absolute error (E_{abs}) is introduced as follows.

$$E_{abs} = T^{'} - T \tag{5}$$

T' − Computing skin temperature by the algorithm
T − Measured skin temperature

Image high frequency noise, such as salt-and-pepper noise, was generated because of lighting intensity and camera senor temperature when images were captured by camera with charge-coupled device. To guarantee accuracy, median filter method was used which belonged to order statistics filter. The median filter had excellent de-noising capability for some random noises, especially salt-and-pepper noise [23].

Program for video magnification, published in the website of Massachusetts Institute of Technology (http://people.csail.mit.edu/mrub/vidmag/), was used with fine-tuning parameters. Noise was also magnified when valid information in the video was magnified. The median filter was used again to guarantee accuracy. Six color parameters from both RGB space and HSV space were analyzed. Information was saved in a numerical matrix after extraction of color parameters from one color space. Skin color saturation, which described color purity and signal intensity in images, had strong correlation with skin temperature.

Table 3
Online non-invasive measuring algorithm based on video magnification.

Algorithm: Online non-invasive measuring algorithm based on video magnification
Input: Hand video samples, 16 × 3000 s × 30 frames per second (fps)
Output: 16 personalized ST model
Initialization: Initial values of parameters
First Step: Training layer by layer and model construction
(1) Remove high frequency noise;
(2) Process the video with magnification technology and filtering again;
(3) Search region of interest, extract saturation and other information;
(4) Construct and optimize cost function, construct the ST model ($T = aS + b$);
(5) Compute median value of slopes with all subjects.
Second Step: Supervised learning and model optimization
(1) Personalized data matrix import, video magnification and filtering processing;
(2) Based on gradient optimization, back propagation method was used to top-down fine tuning model parameters;
(3) Get personalized point and ST model.

4. Results

After 10 min thermal stimulus, video of hand backs were recorded for 50 min by 30 fps, which is 90000 frames totally. Resolution is 1280 × 720. Matlab was used to do programming for video magnification. Personal Computer with i7-5500U CPU, 16 GB RAM and 8 GB graphic memory was used. In EVM algorithm, amplification factor is 10, spatial frequency cutoff is 16, chrom attenuation is 0.1, and the low pass filter parameters are 0.4 and 0.05 respectively. Original images and magnified images of hand backs for the 16 subjects were shown in Fig. 3.

The 90000 frames were processed after video magnification. Region of interest was found and information of skin color saturation was extracted. Corresponding to 1 min sampling interval for hand back skin temperature, average value of hand back skin color saturation within 1 min was extracted based on 1800 frames (30 fps × 60 s). Variations of hand back skin color saturations of 16 subjects were shown in Fig. 4. The trend of linear decreasing was demonstrated. Variations of hand back skin temperatures of 16 subjects were shown in Fig. 5. The trends of linear decreasing were similar.

50 discrete points were plotted in Fig. 6, which used hand back skin color saturation as x-axis and hand back skin temperature as y-axis. Variation trends of the 50 discrete points were shown in red line. Linear regression was used and shown in blue line, which is personalized ST model. The median 96.5 (80.5, 178.2) of the 16 slopes was used for partly personalized ST model (Equation (5)). The equation should be called partly personalized ST model because slope median was used. For further tests with larger sample size, median value of slopes may change slightly. Personalized point for each subject was used, which was obtained under normal room temperature (22 °C in this test) without any thermal stimuli. Partly personalized ST model was shown in green line. When compared green line to blue line, Fig. 6 (1, 8, 9, 10, 14) matched excellently and Fig. 6 (2, 3, 5, 7, 12, 16) matched well.

According to the linear regression, the slope vector [a_1, a_2, …, a_{16}] was obtained and the median value is 96.5. The partly personalized ST model is expressed as

$$T = 96.5 \times S + b_i (i = 1, 2, 3, \ldots, 16) \tag{6}$$

Equation (6)is suitable for East Asia young females. i denotes different users and b_i is personalized intercept of different users.

By comparing predicted skin temperatures from personalized ST

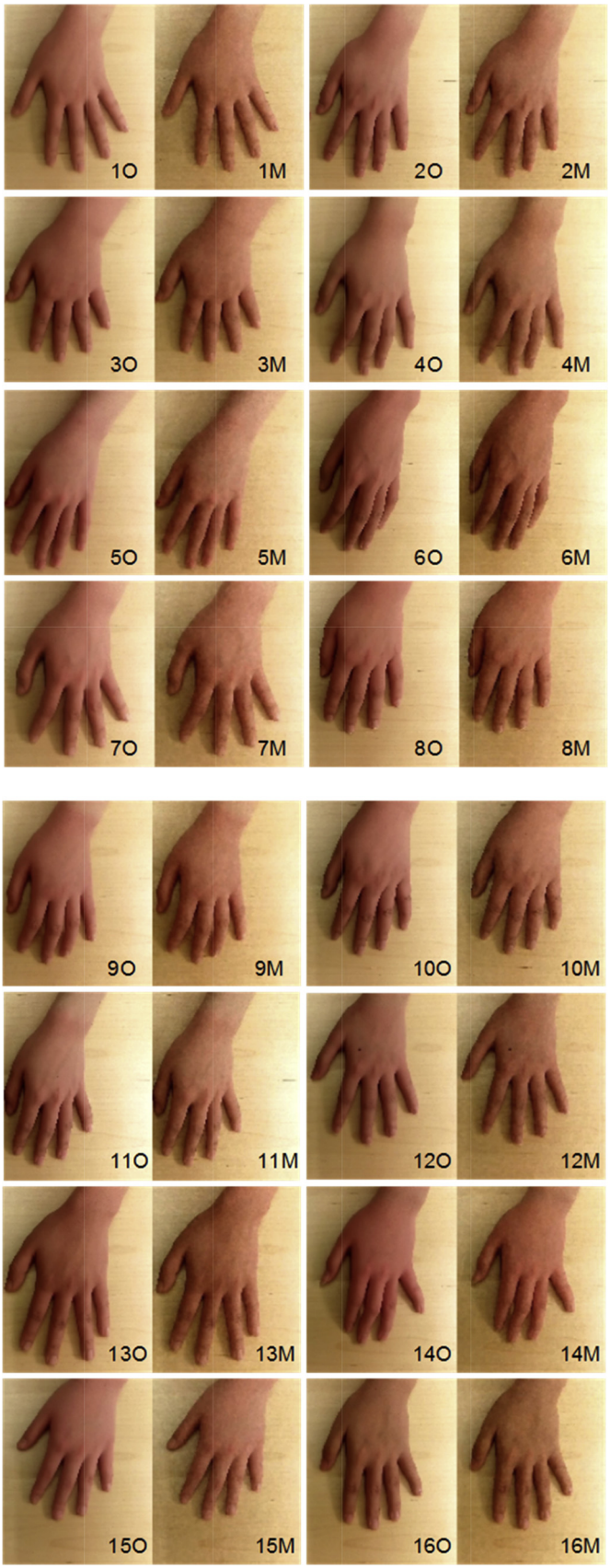


Fig. 3. Hand backs of 16 subjects (O: original images, M: magnified images).

models and measured skin temperatures, distributions of absolute errors for the 16 subjects were shown in Fig. 7. For each subject, first quartile, median and third quartile of the 50 absolute errors were shown. Medians of absolute errors for the 16 subjects changed from −0.10 °C to 0.06 °C. By comparing predicted skin temperatures from partly personalized ST models and measured skin temperatures, distributions of absolute errors for the 16 subjects were shown in Fig. 8. Medians of absolute errors for the 16 subjects changed from −1.32 °C to 0.61 °C. Comparison of standard deviations of absolute errors for the 16 subjects was shown in Fig. 9, which demonstrated the feasibility of the partly personalized ST model.

5. Discussion

In the personalized ST model developed in this paper, the linear relationship between skin color saturation and skin temperature is very obvious. Within the studied cohort (young East Asian females), median of slope for the cohort is close to slopes of most subjects, which should be validated by further study with larger sample size. Young female subjects have relatively delicate skin, which has no skin folds and are sensitive to thermal stimuli. The advantage of personalized ST model is of highly accurate prediction of skin temperature by skin color saturation. The limitation of personalized ST model is that every new subject needs to repeat thermal stimulus process to find accurate linear relationship between skin temperature and skin color saturation, which can be saved in computer database for further use. The advantage of partly personalized ST model is that only basic skin temperature without thermal stimuli and corresponding skin color saturation should be measured, which is much easier for every new subject without repeating thermal stimulus process. The limitation of partly personalized ST model is the prediction accuracy.

Lagrange and Euler method have been developed and used in video magnification field. Lagrange perspective focused on image pixels. Trajectory of pixel motion was analyzed and amplified [13–16]. The method is sensitive to small motions. However, there are two main disadvantages in Lagrange perspective. Firstly, excellent algorithm for estimating pixels trajectory is required, which is not always satisfied. Secondly, due to the attention to pixel level microscopic motion, analysis of the whole image is lacked. When local pixel is enlarged, target object and surrounding background are teared, which need image patching algorithm to compensate. Euler perspective focused on whole image and its region of interest. It analyzed key signals in region of interest of image, including signals in region of interest and highly correlated signals of region of interest. Target signals were amplified [17–20]. The method, which is easy for practical application, started from global image and overcome the limitations of Lagrange method. The situation is similar as that in fluid dynamics field. Lagrange method is used to describe motion of fluid particle. Eulerian method is used to describe flow field at certain time, which is more convenient to establish transportation equations.

Strong thermal stimulus was tried in this pilot study, which caused 4–5 °C temperature changes. Somebody may argue that only weak thermal stimulus with about 2 °C temperature changes happened such as using hand warmer. One reason of performing strong thermal stimulus experiments was to create ST model obviously and easily. For normally happened weak thermal stimuli, further experiments will be conducted. As a new technique, Eulerian video magnification algorithm developed rapidly in the past decade, although it still has some limitations such as noise control. Let's make an analogy. Half a century ago, computational fluid dynamics (CFD) technology was regarded as impossible to simulate airflow environments with complicated geometry

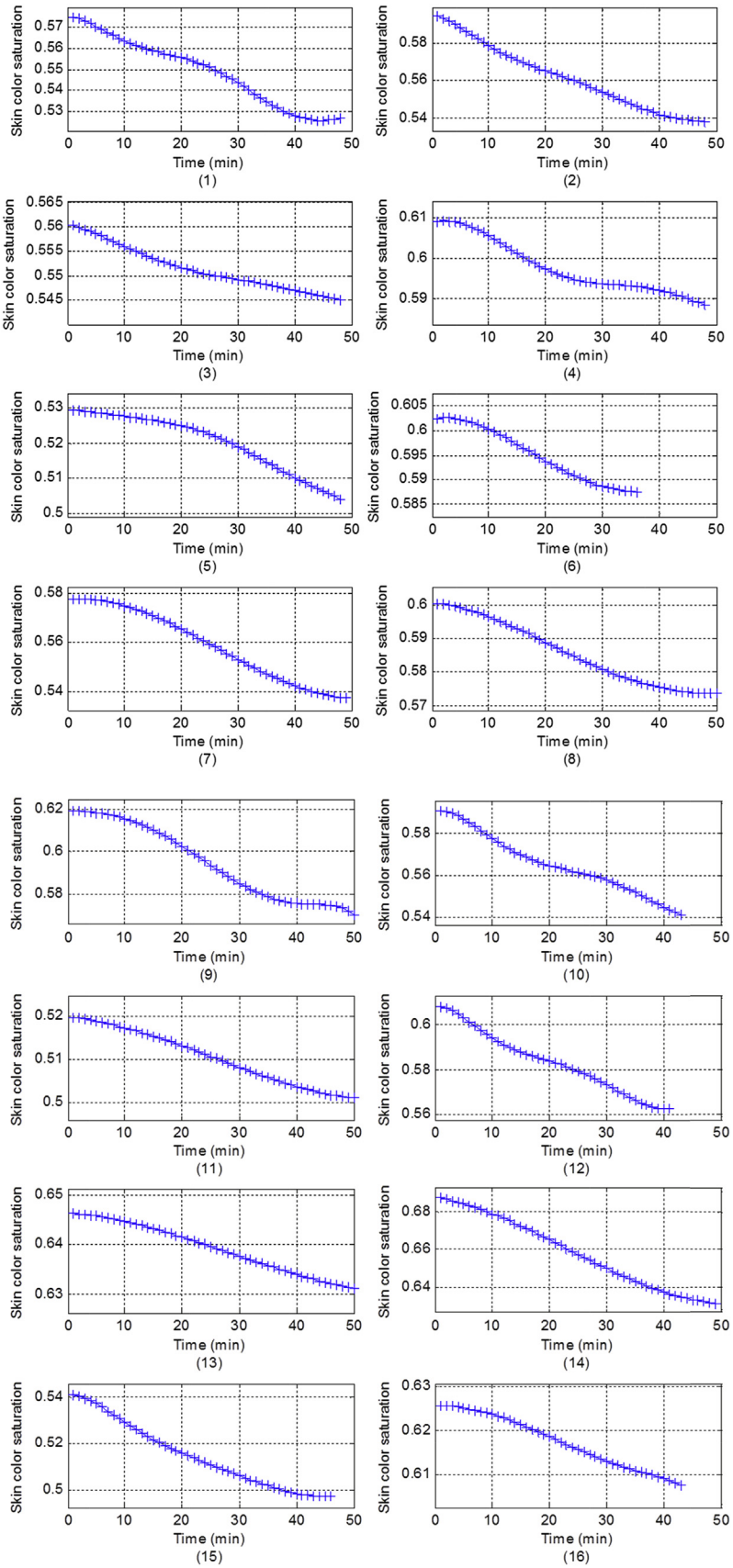


Fig. 4. Variations of hand back skin color saturations of 16 subjects after thermal stimulus.

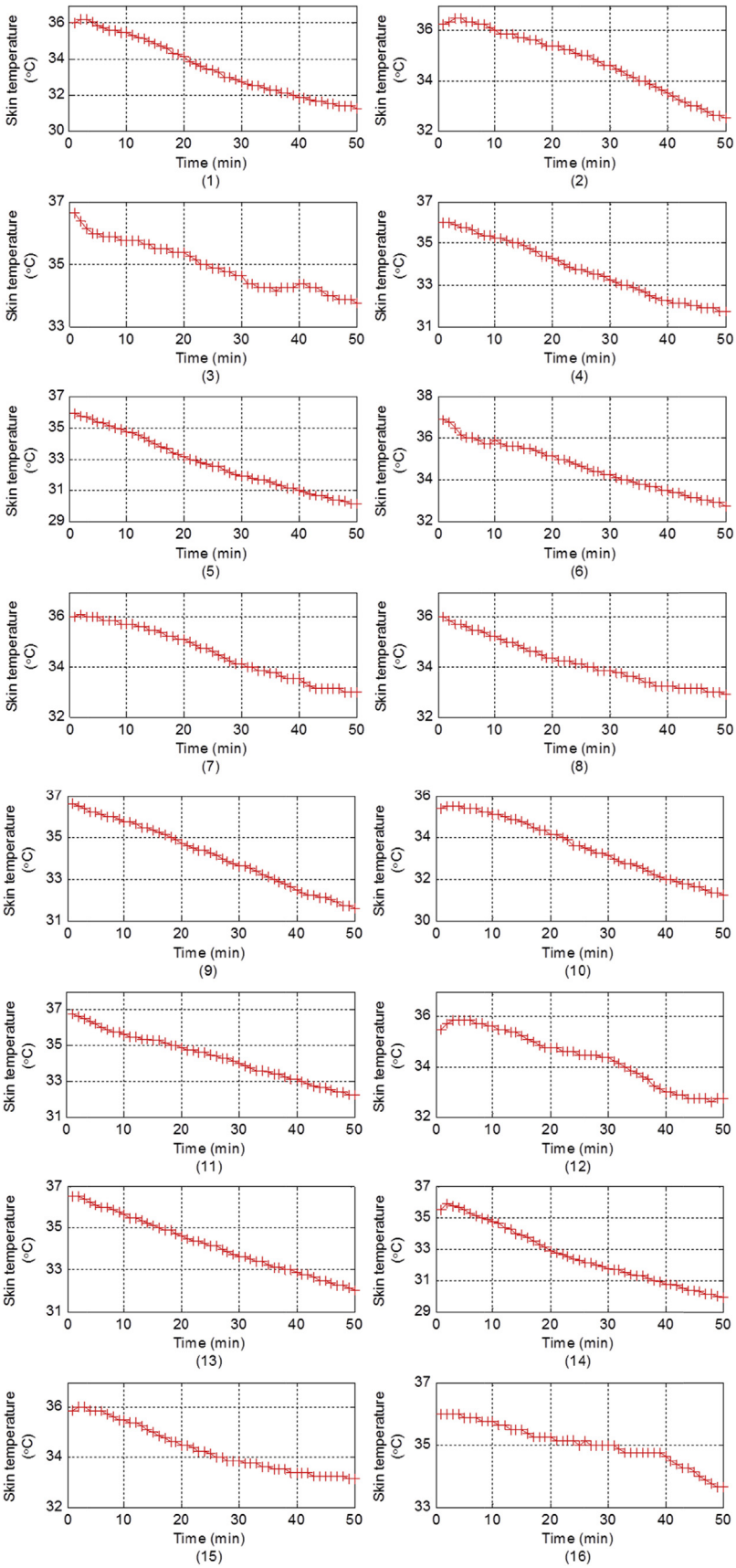


Fig. 5. Variations of hand back skin temperatures of 16 subjects after thermal stimulus.

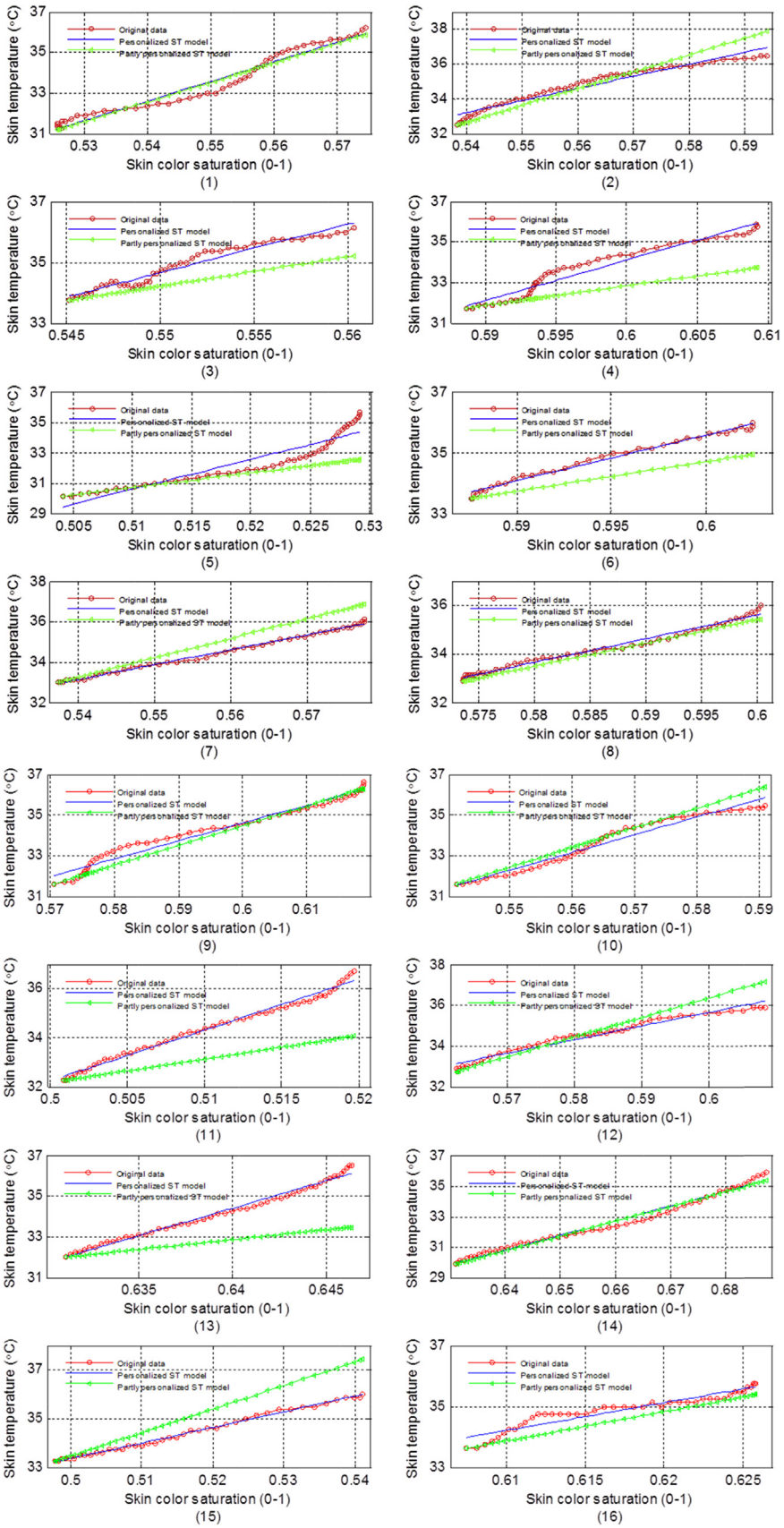


Fig. 6. Regression analysis of skin color saturation and skin temperature.

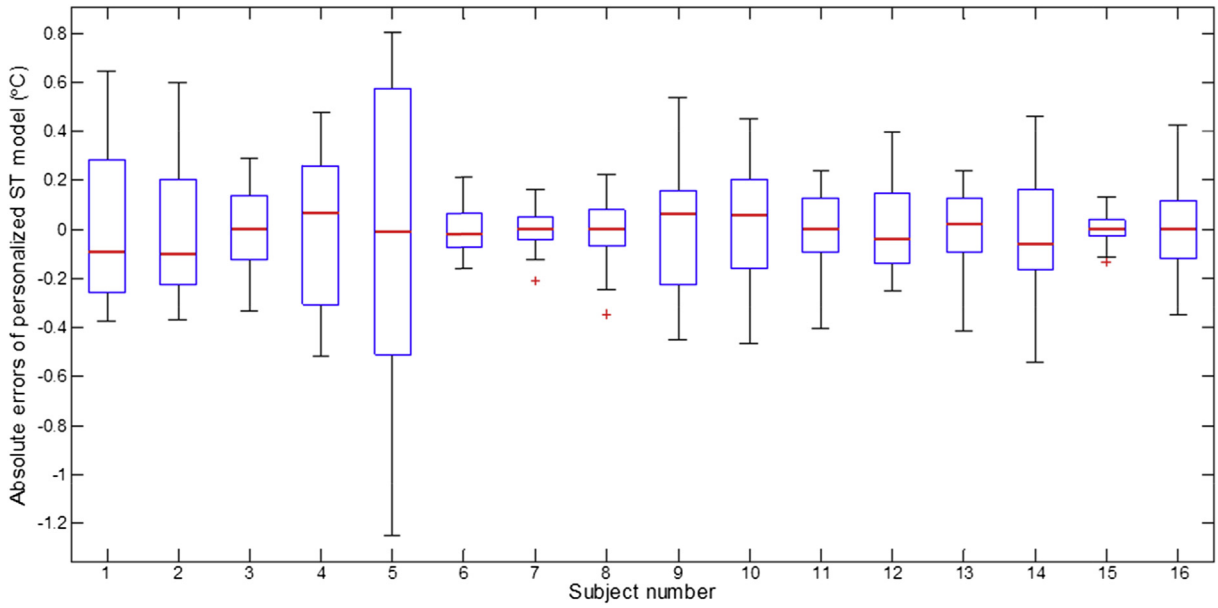


Fig. 7. Absolute errors of personalized ST model (linear regressions) for the 16 subjects.

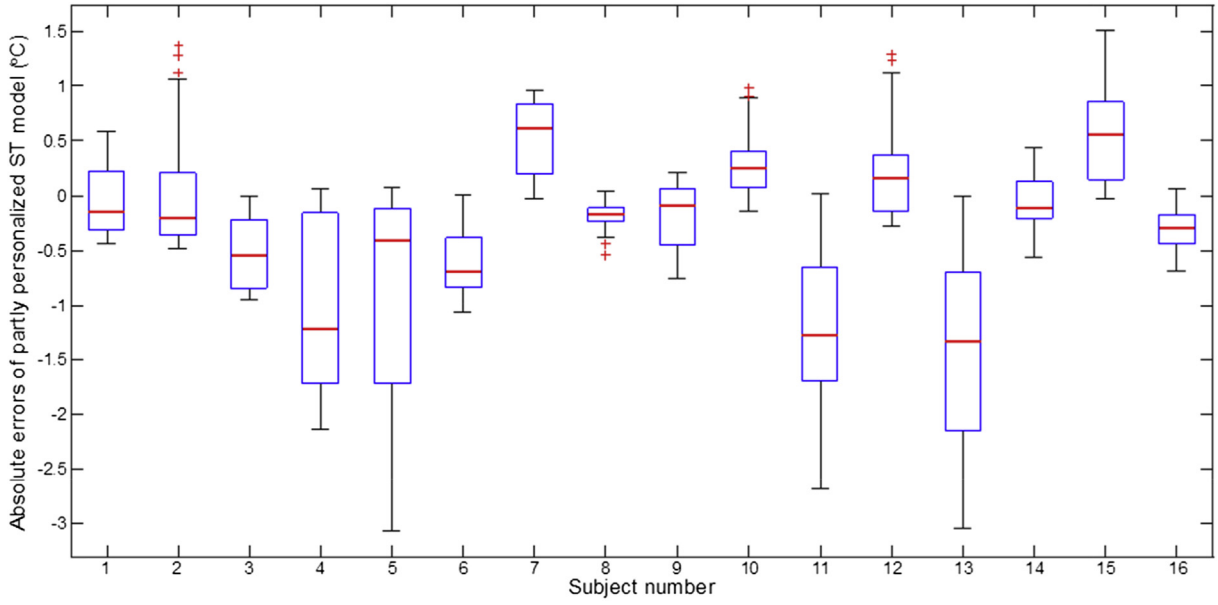


Fig. 8. Absolute errors of partly personalized ST models for the 16 subjects.

because of computational capability limits. However, CFD can be used to finish above mentioned case or even complicated cases within very limited time. Video magnification technique can also be used to measure tiny skin temperature variations in near future.

The skin temperature signal, obtained by video magnification technique, can be used to control not only HVAC systems but also adaptive thermal comfort strategies. For example, occupants will be notified by the signal to adjust personal attires to adapt to the immediate environments. The signal can also be used to control and modulate operable windows.

The original code about EVM was shared by Professor William Freeman from Computer Science and Artificial Intelligence Laboratory of Massachusetts Institute of Technology. It is publicly accessed, which is helpful for validating and generalizing the pilot study in this paper.

6. Conclusions

The main purpose of this study was to find a non-invasive measuring technology for measuring human skin temperature, which can be used as feedback signal for control HVAC systems. The conclusions were drawn as follows.

(1) Eulerian video magnification technology can be used to accurately analyze skin color saturation;
(2) Linear relationship between skin color saturation and skin temperature existed;
(3) Personalized ST model (linear regression) can be used to achieve non-invasive measurements of skin temperatures for young East Asia females with high accuracy and complex

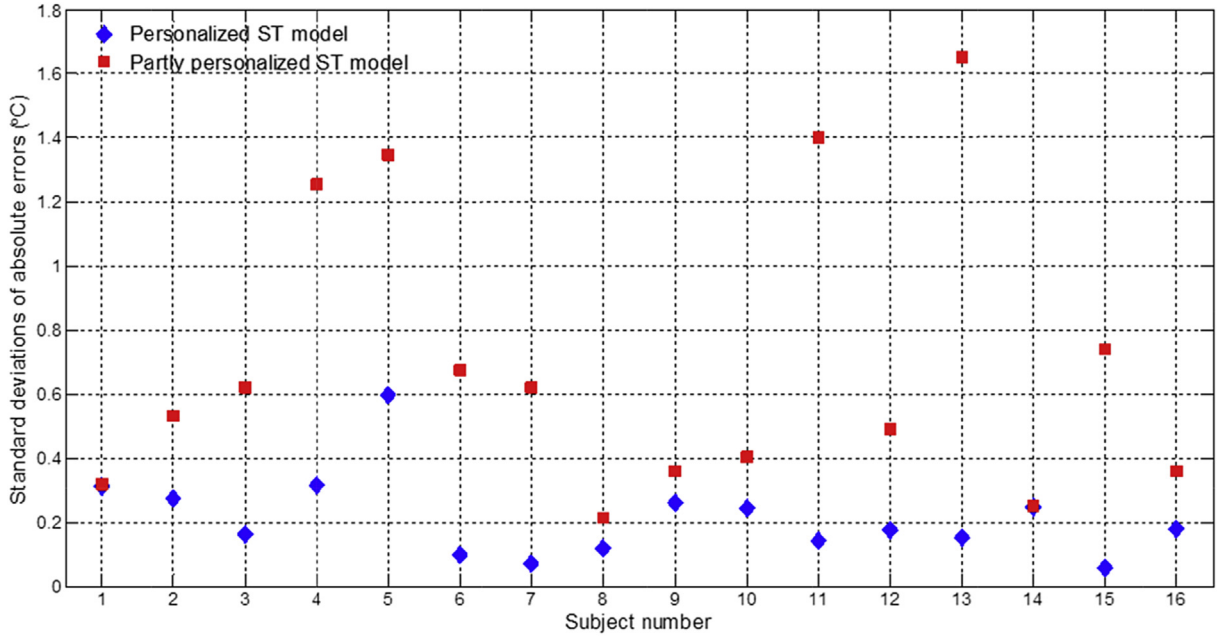


Fig. 9. Comparison of standard deviations of absolute errors between personalized ST models (linear regressions) and partly personalized ST models.

process. Medians of absolute errors changed from −0.10 °C to 0.06 °C;

(4) Partly personalized ST model ($T = 96.5 \times S + b_i$) can be used to achieve non-invasive measurements of skin temperatures for young East Asia females with less high accuracy and simple process. Medians of absolute errors changed from −1.32 °C to 0.61 °C.

There are some limitations in this pilot study. The median slope, based on individual slopes from 16 subjects, is not close enough to each slope. Test with larger sample size should be performed to test the feasibility of partly personalized ST model. Method of deriving thermal sensation, based on skin temperature obtained by video magnification, should be explored for realizing HVAC control in practice.

Acknowledgements

The project is supported by National Natural Science Foundation of China (No. 61401236), Jiangsu Postdoctoral Science Foundation (1601039B), Key Research Project of Jiangsu Science and Technology Department (No. BE2016001-3). The authors would like to acknowledge Professor William Freeman from Massachusetts Institute of Technology for sharing his code about Eulerian Video Magnification.

References

[1] U. EIA, Annual Energy Outlook 2013, US Energy Information Administration, Washington, DC, 2013.
[2] EMA, Singapore Energy Statistics 2014, Energy Market Authority, Singapore, 2014.
[3] X. Yuan, Q. Chen, L.R. Glicksman, A critical review of displacement ventilation, ASHRAE Trans. 104 (1998) 78–90.
[4] Z. Lin, T.T. Chow, C.F. Tsang, Stratum ventilation? A conceptual introduction, in: Proceedings of the 10th International Conference on Indoor Air Quality and Climate (Indoor Air 2005), 2005, pp. 3260–3264.
[5] P.O. Fanger, Human requirements in future air-conditioned environments, Int. J. Refrig. 24 (2001) 148–153.
[6] A.K. Melikov, Personalized ventilation, Indoor Air 14 (2004) 157–167.
[7] F. Bauman, A. Daly, Underfloor Air Distribution (UFAD) Design Guide ASHRAE, Altanta, GA, 2003.
[8] H. Zhang, E. Arens, Y. Zhai, A review of the corrective power of personal comfort systems in non-neutral ambient environments, Build. Environ. 91 (2015) 15–41.
[9] B. Yang, S. Schiavon, C. Sekhar, D. Cheong, K.W. Tham, W.W. Nazaroff, Performance evaluation of an energy efficient stand cooling fan, Build. Environ. 85 (2015) 196–204.
[10] S. Schiavon, B. Yang, Y. Donner, W.C. Chang, W.W. Nazaroff, Thermal comfort, perceived air quality and cognitive performance when personally controlled air movement is used by tropically acclimatized persons, Indoor Air (2016), http://dx.doi.org/10.1111/ina.12352.
[11] H. Zhang, Human Thermal Sensation and Comfort in Transient and Non-uniform Thermal Environments, PhD thesis, University of California, Berkeley, 2003.
[12] A. Ghahramani, G. Castro, B. Becerik-Gerber, X. Yu, Infrared thermography of human face for monitoring thermoregulation performance and estimating personal thermal comfort, Build. Environ. 109 (2016) 1–11.
[13] C. Liu, A. Torralba, W.T. Freeman, F. Durand, E.H. Adelson, Motion magnification, ACM Trans. Graph. 24 (2005) 519–526.
[14] J. Wang, S.M. Drucker, M. Agrawala, M.F. Cohen, The cartoon animation filter, ACM Trans. Graph. 25 (2006) 1169–1173.
[15] M.-Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Opt. Express 18 (2010) 10762–10774.
[16] PHILIPS. Philips vitals signs camera. http://www.vitalsignscamera.com. 2011-08-16/2017-01-20.
[17] M. Fuchs, T. Chen, O. Wang, R. Raskar, H.P. Seidel, H.P.A. Lensch, Real-time temporal shaping of high-speed video streams, Comput. Graph. 34 (2010) 575–584.
[18] H.Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. 31 (2012) 1–8.
[19] M. Rubinstein, N. Wadhwa, F. Durand, W.T. Freeman, H.Y. Wu, Revealing invisible changes in the world, Science 339 (2013) 518–519.
[20] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, Phase-based video motion processing, ACM Trans. Graph. 32 (2013) 1–9.
[21] M. Elgharib, M. Hefeeda, F. Durand, W.T. Freeman, Video magnification in presence of large motions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), 2015, pp. 4119–4127.
[22] A.R. Smith, Color gamut transform Pairs, in: Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 78), 1978, pp. 12–19.
[23] R.C. Gonzalez, R.E. Woods, Digital Image Processing, ISBN-13 978-0131687288, ISBN-10 013168728X, 3.5.3, third ed., Pearson Education, Inc., the United States of America, 2008, pp. 133–136.

Active Tracking Using Color Silhouettes for Indoor Surveillance

Guijin Tang[1,2], Xiaohua Liu[1], Changhong Chen[1,2], Lei Wang[1], Ziguan Cui[1], Zongliang Gan[1], Feng Liu[1], Suhuai Luo[3]
[1]Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
[3]University of Newcastle, NSW 2287, Australia
E-mail: tanggj@njupt.edu.cn

***Abstract*—Pan-Tilt-Zoom (PTZ) cameras play an important role in surveillance systems. In this paper, we propose an active tracker using color silhouettes with a single camera. We firstly apply dilation and erosion operators of morphology to binary difference image to get color silhouettes. We also record the color silhouette of the target which we are interested in. Secondly, we measure the similarity of color silhouette between the observation and the candidates of silhouettes. We exploit the most similar one to update that of the tracked target. Finally, we control the PTZ camera to move according to the location of the tracked target. The experimental results show that the proposed algorithm can effectively track people even though she/he is fully occluded.**

Keywords—active tracking; color silhuette; indoor surveillance;

I. INTRODUCTION

Moving object tracking in videos has been an active research for decades. It is motivated by numerous applications, such as surveillance, traffic control, and virtual reality. However, there are still many challenges in these areas, which make the performance of existing methods far from being desired in solving the emerging real-world problems. These challenges are arisen by the scene change, non-rigid object structures and inter-object occlusion.

Traditional surveillance systems use fixed cameras to track objects, so they only passively receive the information of cameras. In order to expand the camera's view, Pan-Tilt-Zoom (PTZ) cameras are adopted which can track targets automatically. In the field of camera's view, the PTZ will detect the moving target as soon as it appears. Then the PTZ camera will track the active target according to the orientation of the target, and always keep it in the central area of an image. Thus, an active monocular video surveillance system achieves wider coverage as compared to the ones with single fixed vision camera. Moreover, in such cases the system can have a better look at the interesting target, which may be an unusual intrusion activity.

Some works on active tracking have been reported. In [1], Yilmaz *et al.* proposed an algorithm of contour-based people tracking. The algorithm had two major components related to the visual features and the object shape. It evolved the contour from frame to frame by minimizing the energy functional. It could track the complete region of the non-rigid objects and recover occluded object parts. But sometimes occlusion detection of this work would go wrong. Varcheie *et al.*[2] designed a network-based PTZ camera system which could track the human upper body in an online application. However, it would lose the target if the person suddenly changed his motion direction. Chen *et al.*[3] applied multiple cameras in automated surveillance systems. They proposed a novel mapping algorithm that could derive the relative positioning and orientation between two PTZ cameras based on a unified polynomial model without prior knowledge of camera intrinsic parameters. In the work of [4], Elder *et al.* used two cameras. One was a fixed, pre-attentive, low-resolution wide-field camera for detection, while the other was a shiftable, attentive, high-resolution narrow-field camera for confirmation and further analysis. The advantage of this system was a wide FOV (Field of View), but it relied on a communication feedback between two cameras.

In this paper, we focus on the active tracking for indoor surveillance with a single PTZ camera. When the system starts, a counter of moving objects which are detected by temporal differencing is set up. When several successive frames detect one target whose area is bigger than the threshold in each of them, the current moving target will be regarded as the tracked one. At the same time, we record the color feature of the object's silhouette. As the system goes, the color feature is updated if the degree of similarity between the current color silhouette and the latest color silhouette exceeds the similarity threshold. When multi-objects appear in the scenes, we utilize color silhouettes to identify the tracked target. When the object is moving out of the central area of the image, the system will drive the camera to capture the tracked target again.

The rest of the paper is organized as follows. In section II the system overview and framework will be described. Section III formally presents our algorithm. In the subsequent section IV, some experimental results of active tracking with a PTZ camera is discussed and analyzed. Finally, we conclude the paper by highlighting the achievements and future work in section V.

This research was supported in part by China Scholarship Council, the National Nature Science Foundation of China(61172118, 61201164, 61471201, 61501260), the Key Project of Natural Science Research of Jiangsu Higher Education Institutions(13KJA510004), the Natural Science Foundation of Jiangsu Province(BK20130867), the Open Project of State Key Laboratory of Novel Software Technology(Nanjing University)(KFKT2014B10, KFKT2015B24), the Natural Science Foundation of NUPT(NY214039), and the "1311" Talent Plan of NUPT.

978-1-4673-7687-7/15/$31.00 ©2015 IEEE

II. System Overview and Framework

The mobile sensor system is made up of three main components: a PTZ camera, a video capture card, and a computer which will run our tracking method. The framework of the system is shown in Fig. 1.

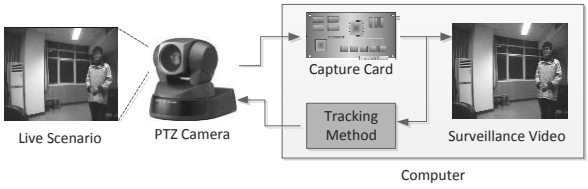


Fig. 1 The framework of the system

In our system, the PTZ camera is Sony EVI D100P whose CCD (Charge-coupled Device) features 440,000 effective picture elements. The video capture card is 10moons SDK 3000 which can offer YUV2, RGB32, RGB24, and RGB555 video formats. A data line is used to connect DB9 port of the computer and VISCA (Video System Control Architecture) port of the camera, thus the tracking method can send the commands to the camera to control its movement.

III. The Tracking Method

In order to construct a completely automatic tracking system, motion detection and location estimation, which are the prerequisites of moving target tracking, need to be done. In other words, we should let the camera know what is to be tracked and where the target is. After the segmentation of the foreground, the system needs to track the target automatically with tracking algorithm from the next frame. Once the target is detected, the system will judge if the camera should move and what angle the camera should conduct.

Our method of active tracking includes three modules. They are the module of color silhouette extraction, the module of similar measurement of color silhouette, and the module of camera control. The module of silhouette extraction can find the motion targets quickly and accurately as soon as active objects appear. The tracking module aims to identify the tracked target among the candidate objects. The last module will drive the camera in terms of the orientation of the tracked target in current frame.

A. Color Silhouette Extraction

Object detection is usually considered as the first step of video tracking. It provides necessary feature information of targets for analysis. Popular methods for motion detection include background subtraction, optical flow and temporal differencing. An appropriate choice among such methods usually depends on different scenarios. The background subtraction uses the subtraction of the current frame and the background image to provide complete foreground data. Hence it is sensitive to dynamic scenes. Optical-flow-based motion segmentation uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. It can be used to detect independently moving objects even in the presence of camera motion. However, most flow computation methods are computationally complex and very sensitive to noise, and cannot be applied to video streams in real time without specialized hardware. The temporal differencing, which can also be called frame differencing, utilizes subtractions of consecutive frames to establish the foreground. It is still effective when the camera is moving. Meanwhile, it has a significantly less computational cost compared with the optical flow approach. Consequently, we use temporal differencing to detect moving objects.

When we carry out the algorithm of two-frame differencing, we use the threshold T_d to obtain a binary image. If the difference is greater than the threshold T_d, the corresponding pixel is considered to be a part of the moving object; otherwise, it is considered to be the background. Its formula is:

$$D_t(x, y) = \begin{cases} 255 & | f_t(x, y) - f_{t-1}(x, y) | \geq T_d \\ 0 & | f_t(x, y) - f_{t-1}(x, y) | < T_d \end{cases} \quad (1)$$

where $f_t(x, y)$ and $f_{t-1}(x, y)$ are the pixel value at time t and that at time t-1, respectively; $D_t(x, y)$ is the binary difference result.

However, the temporal differencing often results in holes inside a silhouette, irregular shape, or the break of a silhouette. So we conduct dilation operation of morphology to deal with the binary difference image. Here we use a 9×9 square structuring element.

After that, we may obtain several silhouettes of moving objects. But unfortunately, some of these candidate objects are not the real moving targets that we are interested in. For example, small silhouettes may result from noise. In another case, a very big silhouette is probably caused by camera movement. Such silhouettes should be excluded, and the corresponding objects should not be tracked accordingly. Therefore, two thresholds T_s and T_b ($T_b > T_s$) of the area size are utilized to choose the proper silhouettes. Let $S_t(i)$ be the state of the ith silhouette at time t, and $A_t(i)$ be the area size. If $S_t(i)$ is equal to 1, it means that the ith silhouette will be kept, otherwise it will be discarded. It can be formulated as:

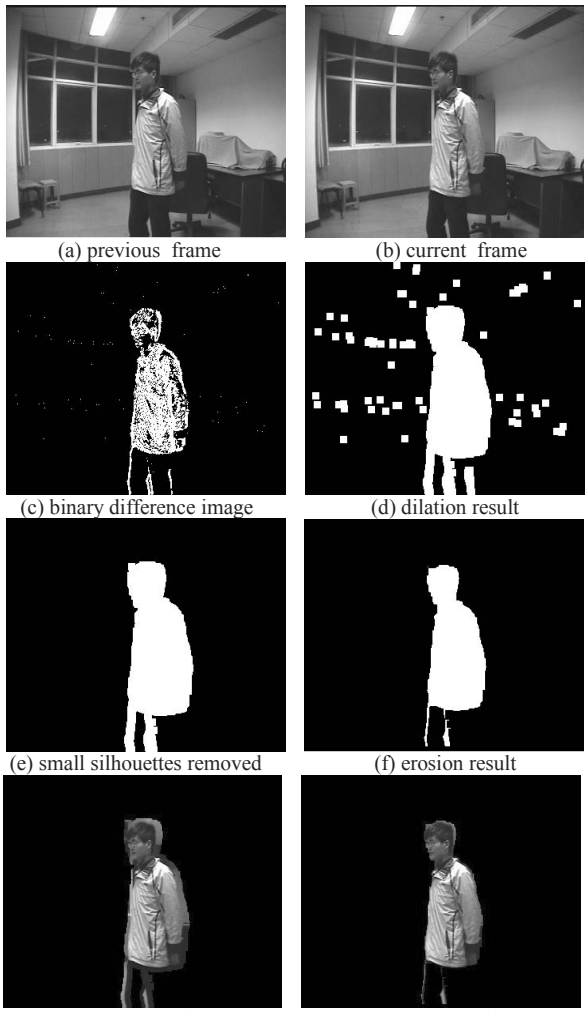
$$S_t(i) = \begin{cases} 1 & T_s \leq A_t(i) \leq T_b \\ 0 & else \end{cases} \quad (2)$$

Now the silhouettes that we are interested in are left. We should make further judgment. Because sometimes the object silhouette still breaks into several parts despite dilation, some silhouette areas of the parts of the moving object can make $S_t(i)$ true. At the same time, we notice that this phenomenon can hardly last for multiple successive frames. That is to say, it happens occasionally. Therefore we set a counter to record the number of moving objects in the scenes when the system starts. If the number of moving objects in the current frame is not equal to the counter, the total objects in the current frame will be discarded. If the number of moving objects is unchanged for certain successive frames, the counter will be updated by this new number.

Because the dilation operation has been conducted during the course of object detection, the silhouette region will be bigger than the actual object. Correspondingly, an erosion operation will be applied to silhouettes.

Then we extract the data in the live image corresponding to the eligible silhouette in the binary image to form a color silhouette. We suppose that the color silhouette can represent the corresponding object.

Fig. 2 shows the process of the color silhouette extraction.



(a) previous frame (b) current frame

(c) binary difference image (d) dilation result

(e) small silhouettes removed (f) erosion result

(g) extracted color silhouette before erosion (h) extracted color silhouette after erosion

Fig. 2. The color silhouette extraction

B. Silhouette Measurement

Now we will track the moving target using color silhouette. In order to match the color silhouettes, we measure the similarity of two color silhouettes according to their major color spectrum histogram (MCSH) [5]. We follow these steps below to obtain the MCSH of color silhouette.

(1)Step1: the normalized geometric distance between two color pixels in the RGB space can be given by the following equation:

$$d(C_p, C_q) = \frac{\|C_p - C_q\|}{\|C_p\| + \|C_q\|} = \frac{\sqrt{(R_p - R_q)^2 + (G_p - G_q)^2 + (B_p - B_q)^2}}{\sqrt{R_p^2 + G_p^2 + B_p^2} + \sqrt{R_q^2 + G_q^2 + B_q^2}} \quad (3)$$

where $C_p=(R_p, G_p, B_p)$ and $C_q=(R_q, G_q, B_q)$ are the color vectors.

(2)Step2: we scan the pixels in the color silhouette to cluster them. The scan direction is left-to-right along the rows and then top-to-bottom. We will regard the first pixel as the center of the first cluster. If the distance between the subsequent pixel and the center of an existing cluster is shorter than the threshold, the current pixel will belong to the cluster, and the element number of the cluster will add one. Otherwise it will be the center of a new cluster.

(3)Step3: After all pixels are traversed, K-means clustering method is used to refine the cluster centers. The objects' pixels are scanned in row-major order. We compute the distance between the current pixel and each cluster center, and assign the pixel to the closest cluster. Then, the center of this cluster is updated. The update strategy is defined as:

$$\begin{cases} R_c(i) = w(i)R(i) + (1 - w(i))R_c(i-1) \\ G_c(i) = w(i)G(i) + (1 - w(i))G_c(i-1) \\ B_c(i) = w(i)B(i) + (1 - w(i))B_c(i-1) \end{cases} \quad (4)$$

where i is the current number of pixels in the cluster, $R(i)$, $G(i)$, $B(i)$ are the RGB components of the ith pixel, $R_c(i)$, $G_c(i)$, $B_c(i)$ are the center of clustered pixels after the ith pixel has been processed, and $w(i)$ denotes the coefficient which is equal to $1/i$.

So a MCSH representation can be obtained for a color silhouette. Fig. 3 shows the MCSH of Fig. 2(h). The horizontal axis and vertical axis denote the color components and the frequency of color components, respectively.

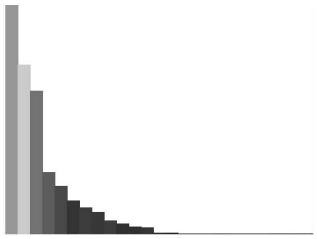


Fig. 3. The MCSH of Fig.2(h)

Then we can exploit the method of [5] to measure the similarity of two silhouettes.

If the value of silhouette similarity between a candidate object and the recorded true object is bigger than the threshold, the current candidate object will be regarded as the true target. Consequently, the MCSH representation of the recorded true target of the system will be updated by that of the just selected true one.

C. Camera Control

In our system, the EVI-D100P that serves as the peripheral device is connected to the computer using communication conforming to the RS-232C standard with 9600 baud, 8 data bits, 1 start bit, 1 stop bit, and no parity.

The basic unit of VISCA communication is called a packet (Fig. 4). The first byte of the packet is called the header which comprises the sender's and receiver's addresses. The header of the packet sent to the EVI-D100P assigned address 1 from the computer (address 0) is hexadecimal 81H. The next part is the message whose length varies from 1 byte to 14 bytes. Every message has different meaning which can control the camera to

conduct the corresponding operation. The end of the packet is the terminator which means the total packet has been transmitted.

Fig. 4. Packet structure

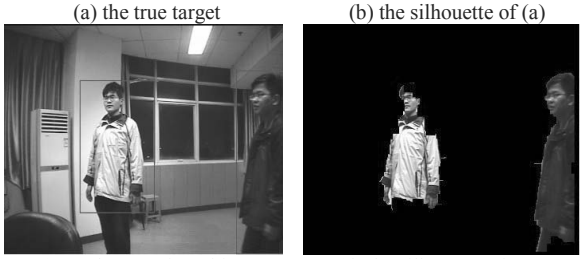
The tracked target is always adjusted in the middle of the image in an active surveillance system. However, in order to keep a smooth video, we should not control the camera to move every time when the tracked target changes a location. If the distance between the centroid location of the target and the center of the image exceeds the threshold, the camera will move.

IV. EXPERIMENTAL RESULTS

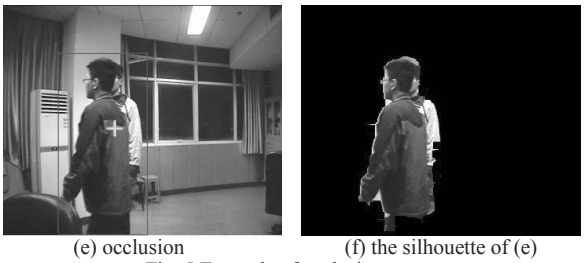
In an active tracking system with a single PTZ camera, there is always only one true target. We consider the first moving object whose silhouette meets the above conditions to be a true target, and call the other moving objects false targets. In order to test the effect of our method, we use real-life video sequences of indoor scenario. The tracking algorithm has been tested over events such as entering or leaving the FOV of the camera and occlusion with other people in the scene. By trial and error, it is found that the values of $T_d = 8$, $T_s = 0.065*W*H$, and $T_b = 0.4*W*H$ can work well in our experiments, while the parameters of W and H mean the image width and height, respectively.

Fig. 5 shows the tracking results when an occlusion occurs. The red box denotes the detected moving object, and the green cross means the location which records the centroid of the true target in the current frame. We can see that the tracking for the true target will not be affected by false objects in our system. When there are several candidates, our algorithm can identify the true target. When the occlusion occurs, the overlapping silhouette is not similar with that of the true target. So our algorithm still keeps the latest correct location of the true target.

(a) the true target (b) the silhouette of (a)



(c) another moving object (d) the silhouette of (c)



(e) occlusion (f) the silhouette of (e)

Fig. 5 Example of occlusion

Fig. 6 illustrates the movement of the PTZ camera when the tracked target will leave the camera's FOV. Our algorithm will change the camera angle to make the tracked target in the central area of an image.

(a) the tracked object near the edge (b) keep the tracked object in the center

Fig. 6 Example of leaving the FOV

V. CONCLUSION

In this paper, we propose an approach to track a target actively with a PTZ camera. We exploit the color silhouette which is represented by MCSH as the feature of the tracked target. In order to get the complete and accurate silhouette, the dilation and erosion operations are applied to the binary difference image. The experimental results show that our system can work well. In the future, more features will be considered in our approach.

REFERENCES

[1] A.Yilmaz, X.Li, and M.Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Issue 11, 2004, pp. 1531-1536.

[2] P.D.Z.Varcheie and G.-A.Bilodeau, "People tracking using a network-based PTZ camera," Machine Vision and Applications, Volume 22, Issue 4, July 2011, pp.671-690.

[3] C.C. Chen, Y.Yao, A. Drira, A. Koschan and M. Abidi, "Cooperative mapping of multiple PTZ cameras in automated surveillance systems," IEEE Conference on Computer Vision and Pattern Recognition, June 2009, pp.l078-1084.

[4] J. H. Elder, S.J.D. Prince, Y. Hou, M. Sizintsev, and E. Olevsky, "Pre-attentive and attentive detection of humans in wide-field scenes," International Journal of Computer Vision, Volume 72, Issue 1, 2007, pp. 47-66.

[5] C.Madden, E.D.Cheng, and M.Piccardi, "Tracking people across disjoint camera views by an illumination-tolerant appearance representation," Machine Vision and Applications, Volume 18, Issue 3, May 2007, pp.233-247.

[6] H.H.Yeh, J.Y.Chen, C.R.Huang, and C.S.Chen, "An adaptive approach for overlapping people tracking based on foreground silhouettes," IEEE 17th International Conference on Image Processing, 2010, pp.3489-3492.

[7] C.Liu, R.Cao, S.Jia, Y. Zhang, B.Wang, Q.Zhao, "The PTZ tracking algorithms evaluation virtual platform system," International Conference on Multisensor Fusion and Information Integration for Intelligent Systems, 2014, pp.1-6.

[8] M.Kamaraj and Balakrishnan, "An improved motion detection and tracking of active blob for video surveillance," Fourth International Conference on Computing, Communications and Networking Technologies, 2013, pp.1-7.

[9] J.Satake and J.Miura, "Stereo-based multi-person tracking using overlapping silhouette templates," 20th International Conference on Pattern Recognition, 2010, pp.4304-4307.

[10] J.Schiel and R.Green, "Adaptive human silhouette extraction with chromatic distortion and contour tracking," 28th International Conference of Image and Vision Computing New Zealand, 2013, pp.288-292.

[11] S.Saraceni, A.Claudi, and A.F.Dragoni, “An active monitoring system for real-time face-tracking based on mobile sensors,” 54th International Symposium ELMAR, September 2012, pp.53-56.

[12] B.J.Lee and K.M.Yi, “Scale preserving PTZ tracking with size estimation using tilt sensory data,”11th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2014, pp.289-294.

[13] D.Comaniciu, V.Ramesh, and P.Meer, "Kernel-based object tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume.25, Issue 5, 2003, pp.564-577.

Convolutional Neural Networks and Feature fusion for Bimodal Emotion Recognition on the EmotiW 2016 Challenge

Jingjie Yan*, Bojie Yan†, Guanming Lu Qinyu Xu Haibo Li Xiaogang Cheng*, Xia Cai‡
*College of Telecommunications and Information Engineering Nanjing University of Posts and Telecommunications, China
†Department of Geography Minjiang University,China
‡Shanghai Technical Institute of Electronics and Information, China

***Abstract*—In the emotion recognition area, it is a more difficult task for recognizing the emotion data from movie or other spontaneous scenes compared to those laboratory scenes. On the base of the AFEW 6.0 database, we present a bimodal emotion recognition approach using the convolutional neural networks and feature fusion method. Firstly, we cut out the facial images and get the audio emotion data from videos respectively. Then, the convolutional neural networks method, the Gabor method and the openSMILE tool are adopted to extract the corresponding features, and three fusion methods including the principal component analysis (PCA) fusion, kernel cross-modal factor analysis (KCFA) fusion and the sparse kernel reduced-rank regression (SKRRR) fusion are utilized to integrate the forgoing facial feature and audio feature in the feature level. At last, the results on the AFEW 6.0 database show that the accuracy rate of the PCA fusion method and the SKRRR fusion method are 53.46% and 50.93% with the svm classifier respectively and are higher than the baseline of the EmotiW 2016 whose accuracy rate is 40.47%.**

***Index Terms*—Bimodal emotion recognition, Convolutional Neural Networks, EmotiW 2016 Challenge, Feature fusion**

I. Introduction

Emotion recognition and analysis based on the video or audio data is a research hotspot in computer vision and affective computing domain among the latest several tens of years [1], [2], [3], [4], [5]. Along with the in-depth research and implementation of emotion recognition approach, the main attention is diverted from the laboratory scenes to realistic or spontaneous scenes. The facial expression data or audio data obtained from the laboratory scenes are relative simple, clear and apparent, but they have a certain gap with those emotion data on the realistic or spontaneous scenes. Therefore, the emotion method learned from the data of laboratory scenes may be unfit for the realistic scene [1], [6], [7]. Moreover, it is obvious that the task of recognizing the emotion data from movie or those spontaneous scenes is more difficult than the laboratory scenes [6], [8].

In last several years, in order to carry forward the research of emotion recognition in the realistic or spontaneous scenes, some researches conduct a series of challenges including the Facial Expression Recognition and Analysis (FERA) [9], Audio Video Emotion Challenge (AVEC) [10], [11] and Emotion Recognition in the Wild (EmotiW)[6], [12], [13], which are based on the movie or other spontaneous scenes [1], [7], [8]. The above-mentioned three emotion recognition challenges are based on the monomodal or multimodal such as the facial expression and speech modality.

In EmotiW 2014, Liu et al. [7] adopt three image feature representation methods including HOG, Dense SIFT and CNN and kernel-based representation modal to conduct bimodal emotion recognition, and their best classification rate reaches up to 50.4%. Sun et al. [14] conduct bimodal emotion recognition based on the the decision level fusion strategy by extracting four classic facial expression features including LPQ-TOP, PHOG, SIFT and LBP-TOP and speech feature. Chen et al. [15] present a bimodal emotion recognition approach based on HOG-TOP and multiple kernel learning, and the best recognition rate of this approach is 50.4%.

In EmotiW 2015, Zong et al. [1] only utilize the facial expression modality and conduct the cross-domain facial expression recognition based on the transductive transfer linear discriminant analysis approach, and their best classification rate is 50%. Yu et al. [8] also only use the facial expression modality and adopt the fusion of multiple deep CNN to extract facial expression feature, and their best recognition rate reaches up to 61.29%. Ng et al. [16] perform facial expression recognition based on the CNN modal and its best classification rate reaches up to 55.6%. Levi et al. [17] integrate the LBP method and the CNN modal to conduct facial expression recognition and get the recognition rate of 54.56%. Cruz et al. [18] fuse a variety of facial expression features and multiple speech features to conduct bimodal emotion recognition. Kahou et al. [19] utilize two fusion strategy to perform bimodal emotion recognition by integrating two deep learning modal including CNN and Recurrent neural networks (RNN), and its best recognition rate reaches up to about 52.88%.

In EmotiW 2016, there are two different emotion recognition challenges including the traditional video-based emotion recognition and the new group level emotion recognition [1], [6], [20], [21], [22]. For the video-based emotion recognition challenges, we present a bimodal emotion recognition approach using the convolutional neural networks modal (CNN) [7], [8], [16], [23], [24], [25], [26], [27], [28] and feature fusion approach. Firstly, we separate the audio emotion data and cut out the valid facial images from the videos of AFEW 6.0 database [6], [20], [21], [22]. Then, the CNN method, the

978-1-5386-1937-7/17/$31.00 ©2017 IEEE

Gabor method [29], [30] and the Opensmile tool [4], [31] are adopted to extract the corresponding features from the previously received audio emotion data and facial images. At last, three fusion methods including the principal component analysis (PCA) fusion [32], kernel cross-modal factor analysis (KCFA) fusion [33] and the sparse kernel reduced-rank regression (SKRRR) [4] fusion are utilized to integrate the forgoing facial feature and audio feature in the feature level.

The rest of paper includes the next four parts. Section II introduces the feature extraction method of the CNN modal and the Opensmile tool. Three fusion methods including the PCA fusion, the KCFA fusion and the SKRRR fusion are showed in Section III. Section IV describes the experiment result of the AFEW 6.0 database. Section V is the conclusion of this paper.

II. Feature Extraction

A. Facial Expression Feature

From the result of the EmotiW 2013, EmotiW 2014 and EmotiW 2015, we can see that many deep learning based methods are used as feature representation method and obtain better recognition result compared to those traditional local feature extraction methods such as SIFT, LBP, LBP-TOP and so on [1], [6], [7], [8], [12], [13], [14], [15], [16], [17], [19]. Among those deep learning based feature representation methods, CNN is one of the most powerful approach and can effectively describe the facial expression emotion information [7], [16], [34]. Therefor, we also adopt the CNN modal to extract facial expression feature from the AFEW 6.0 database in this paper.

According to [7], [16], [24], [25], Alex Net is regarded as one of the most powerful CNN modal and acquire good performance in recognition task. The hierarchical structure, the number of layer and other details of the Alex Net can see [7], [16], [24], [25]. In our experiment, we firstly cut out the valid facial images from the videos of the AFEW 6.0 database and resize those facial images to 256×256, then we adopt two different Alex Net based pattern to extract facial expression feature. The first pattern directly invokes the pretraining Alex Net modal to train the training data and test the val data of the AFEW 6.0 database [6]. After obtaining the optimal modal, it invokes the optimal Alex Net modal to extract the feature of the test data of the AFEW 6.0 database. The first pattern is called as Alex Net-1 in this paper. Similar to the literature of [16], the second pattern firstly utilizes the FER-2013 database [35] to train the pretraining Alex Net modal in view of the small sample of the AFEW 6.0 database, and then conduct the same procedure of the above first pattern to extract the feature of the test data of the AFEW 6.0 database. The second pattern is called as Alex Net-2 in this paper. The details of the procedure and parameter of the Alex Net can see [16]. Moreover, we also extract the Gabor feature except the CNN feature and integrate the Gabor feature and the CNN feature as the facial expression feature.

B. Speech Feature

In the last three EmotiW challenges (EmotiW 2013, EmotiW 2014 and EmotiW 2015), some approaches just take advantage of the facial expression modality and withdraw the audio modality [1], [8], [16], [17]. For those making use of the audio modality approaches, they majority utilize the openSMILE or openEAR [36] tool to extract the feature of the audio modality [4], [6], [7], [14], [15], [19], [37]. In this paper, after separating the audio emotion data from the videos of AFEW 6.0 database, we firstly get rid of the silence part of each audio and then similar to [6], we make use of the openSMILE tool to extract features of the audio modality. The detailed composition of the extracted audio feature can see the literature of [6].

III. Feature Fusion Method

According to the result of last three EmotiW challenges, we can find that most approaches use the decision level fusion and seldom utilize the feature level fusion method [4], [7], [14], [15], [18], [19], [32], [38], [39], [40]. In this paper, three feature fusion methods including the PCA fusion, the KCFA fusion and the SKRRR fusion are respectively utilized to integrate the forgoing facial feature and audio feature in the feature level.

A. PCA Fusion

Suppose the facial expression feature which integrating the Gabor feature and the CNN feature is $\mathbf{X}$ and the audio feature is $\mathbf{Y}$. Then the PCA fusion method can be represented as [4], [32], [39], [41]

$$\mathbf{Fusion}_{PCA} = \begin{pmatrix} \mathbf{M}^T\mathbf{X} \\ \mathbf{N}^T\mathbf{Y} \end{pmatrix}. \tag{1}$$

where $\mathbf{M}$ and $\mathbf{N}$ respectively represent the corresponding projection matrix of the PCA approach.

B. KCFA Fusion

According to [33], the KCFA fusion method can be represented as

$$\begin{aligned} \arg\min_{\mathbf{M},\mathbf{N}} \quad & \|\mathbf{M}^T\varphi(\mathbf{X}) - \mathbf{N}^T\psi(\mathbf{Y})\|_F^2 \\ & s.t. \mathbf{M}\mathbf{M}^T = \mathbf{I}, \mathbf{N}\mathbf{N}^T = \mathbf{I}, \end{aligned} \tag{2}$$

then it can be rewritten as

$$\begin{aligned} \arg\min_{\mathbf{M},\mathbf{N}} \quad & tr\{\varphi(\mathbf{X})^T\varphi(\mathbf{X})\} + tr\{\psi(\mathbf{Y})^T\psi(\mathbf{Y})\} \\ & -2tr\{\varphi(\mathbf{X})^T\mathbf{M}\mathbf{N}^T\psi(Y)\} \\ & s.t. \mathbf{M}\mathbf{M}^T = \mathbf{I}, \mathbf{N}\mathbf{N}^T = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{M}$ and $\mathbf{N}$ respectively represent the corresponding projection matrix of the KCFA fusion approach. Suppose the kernel matrix $\mathbf{K}_X = \varphi(\mathbf{X})^T\varphi(\mathbf{X})$, $\mathbf{K}_Y = \psi(\mathbf{Y})^T\psi(\mathbf{Y})$, then it can be rewritten as another form

$$\begin{aligned} \arg\min_{\mathbf{M},\mathbf{N}} \quad & tr\{\mathbf{K}_X\} + tr\{\mathbf{K}_Y\} \\ & -2tr\{\varphi(\mathbf{X})^T\mathbf{M}\mathbf{N}^T\psi(Y)\} \\ & s.t. \mathbf{M}\mathbf{M}^T = \mathbf{I}, \mathbf{N}\mathbf{N}^T = \mathbf{I}. \end{aligned} \tag{4}$$

According to [33], (4) is equivalent to

$$\begin{aligned} \arg\max_{\mathbf{M},\mathbf{N}} \quad & tr\{\varphi(\mathbf{X})^T\mathbf{M}\mathbf{N}^T\psi(Y)\} \\ & s.t.\mathbf{M}\mathbf{M}^T=\mathbf{I},\mathbf{N}\mathbf{N}^T=\mathbf{I}. \end{aligned} \tag{5}$$

Finally, the fusion feature of the KCFA method can be represented as

$$\mathbf{Fusion}_{KCFA}=\begin{pmatrix}\mathbf{M}^T\mathbf{K}_X\\ \mathbf{N}^T\mathbf{K}_Y\end{pmatrix}. \tag{6}$$

C. SKRRR Fusion

According to the literature of [4], the SKRRR fusion method can be represented as

$$\begin{aligned} \arg\min_{\mathbf{M},\mathbf{N}} \quad & \|\varphi(\mathbf{X})-\varphi(\mathbf{X})\mathbf{M}\mathbf{N}^T\psi(\mathbf{Y})^T\psi(\mathbf{Y})\|_F^2 \\ & +\lambda\|\mathbf{M}\|_1+\mu\|\mathbf{N}\|_1, \end{aligned} \tag{7}$$

then (7) can be represented as the following kernel matrix form

$$\begin{aligned} & \arg\min_{\mathbf{M},\mathbf{N}} tr\{\varphi(\mathbf{X})^T\varphi(\mathbf{X})\}+ \\ & tr\{\psi(\mathbf{Y})^T\psi(\mathbf{Y})\mathbf{N}\mathbf{M}^T\varphi(\mathbf{X})^T\varphi(\mathbf{X})\mathbf{M}\mathbf{N}^T\psi(\mathbf{Y})^T\psi(\mathbf{Y})\} \\ & -2tr\{\varphi(\mathbf{X})^T\varphi(\mathbf{X})\mathbf{M}\mathbf{N}^T\psi(\mathbf{Y})^T\psi(\mathbf{Y})\} \\ & +\lambda\|\mathbf{M}\|_1+\mu\|\mathbf{N}\|_1 \\ = \quad & \arg\min_{\mathbf{M},\mathbf{N}} tr\{\mathbf{K}_X\}+tr\{\mathbf{K}_Y\mathbf{N}\mathbf{M}^T\mathbf{K}_X\mathbf{M}\mathbf{N}^T\mathbf{K}_Y\} \\ & -2tr\{\mathbf{K}_X\mathbf{M}\mathbf{N}^T\mathbf{K}_Y\}+\lambda\|\mathbf{M}\|_1+\mu\|\mathbf{N}\|_1, \end{aligned} \tag{8}$$

where $\mathbf{M}$ and $\mathbf{N}$ respectively represent the corresponding projection matrix of the SKRRR fusion approach, $\mathbf{K}_X=\varphi(\mathbf{X})^T\varphi(\mathbf{X})$, $\mathbf{K}_Y=\psi(\mathbf{Y})^T\psi(\mathbf{Y})$.

The procedure of the SKRRR algorithm can see the literature of [4]. At last, similar to the above KCFA feature fusion, it can obtain the fusion feature in the form of the following pattern [4], [32], [39]

$$\mathbf{Fusion}_{SKRRR}=\begin{pmatrix}\mathbf{M}^T\mathbf{K}_X\\ \mathbf{N}^T\mathbf{K}_Y\end{pmatrix}. \tag{9}$$

IV. Experiments

In EmotiW 2016, there are two different emotion recognition challenges including the traditional video-based emotion recognition and the new group level emotion recognition [1], [6], [20]. The traditional video-based emotion recognition is similar to the EmotiW 2014 and EmotiW 2015 challenges, and the AFEW 6.0 video database of the EmotiW 2016 is gathered on the base of the AFEW 4.0 video database of the EmotiW 2014 and the AFEW 5.0 video database of the EmotiW 2015. Moreover, the AFEW 6.0 video database also is composed of three subsets including training, validation and testing whose corresponding sample size are 773, 383 and 593 respectively, and it has seven emotion categories including Anger, Sad, Neutral, Surprise, Disgust, Happiness and Fear [1], [4], [6], [7], [8], [16], [20]. Moreover, different with the EmotiW 2014 and EmotiW 2015 challenges, the EmotiW 2016 firstly brings in the TV data except the traditional movie data [20]. A few samples of the AFEW 6.0 video database is shown in the Fig.2

Fig. 1. A few samples of the AFEW 6.0 video database.

The result of the baseline, the audio only, the video only ((Alex Net-1+Gabor) and (Alex Net-2+Gabor)), the PCA fusion method, the KCFA fusion method and the SKRRR fusion method with the SVM classifier on the EmotiW 2016 are given on Table I.

From Table I, we can observe that the accuracy rate of the video based on Alex Net-1 and Gabor or Alex Net-2 and Gabor is better than the baseline which is based on LBP-TOP [20] on the validation subset. The accuracy rate of the audio is all obviously lower than the video based on Alex Net-1 and Gabor, the video based on Alex Net-2 and Gabor and the baseline on the validation subset. Moreover, the accuracy rate of the PCA fusion method, the KCFA fusion method and the SKRRR fusion method are all better than the audio, the video based on Alex Net-1 and Gabor, the video based on Alex Net-2 and Gabor and the baseline on the validation subset. On the testing subset, the accuracy rate of the PCA fusion method and the SKRRR fusion method based on Alex Net-2 and Gabor are 53.46% and 50.93% respectively and are higher then the baseline of the EmotiW 2016 whose accuracy rate is 40.47%.

TABLE I
THE RESULT OF THE BASELINE, THE AUDIO ONLY, THE VIDEO ONLY, THE PCA FUSION METHOD, THE KCFA FUSION METHOD AND THE SKRRR FUSION METHOD WITH THE SVM CLASSIFIER ON THE EMOTIW 2016.

Method	Validation	Testing
Baseline (LBP-TOP) [20]	38.81%	40.47%
Audio	36.03%	
Video-1 (Alex Net-1+Gabor)	45.95%	
Video-2 (Alex Net-2+Gabor)	42.04%	
PCA fusion (Video-1+Audio)	48.30%	46.54%
KCFA fusion (Video-2+Audio)	46.48%	
SKRRR fusion (Video-2+Audio)	44.91%	50.93%
PCA fusion (Video-2+Audio)	48.30%	53.46%

The confusion matrix of the audio and the video based on Alex Net-2 and Gabor on the validation subset are respectively

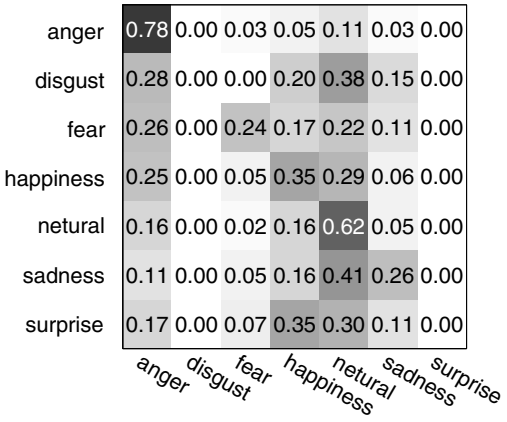


Fig. 2. The confusion matrix of the audio on the validation subset.

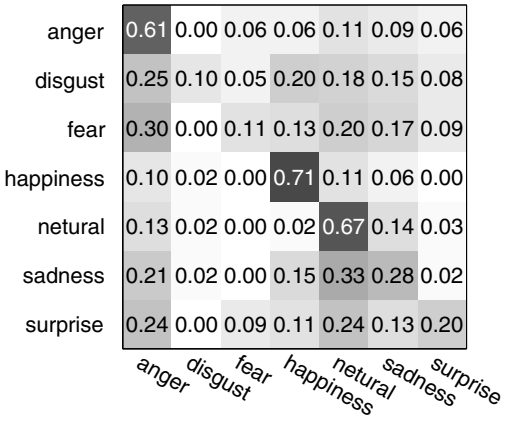


Fig. 3. The confusion matrix of the video based on Alex Net-2 and Gabor on the validation subset.

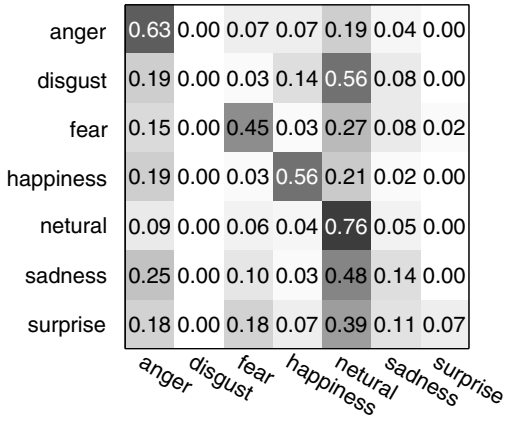


Fig. 4. The confusion matrix of the SKRRR fusion method based on Alex Net-2 and Gabor on the testing subset.

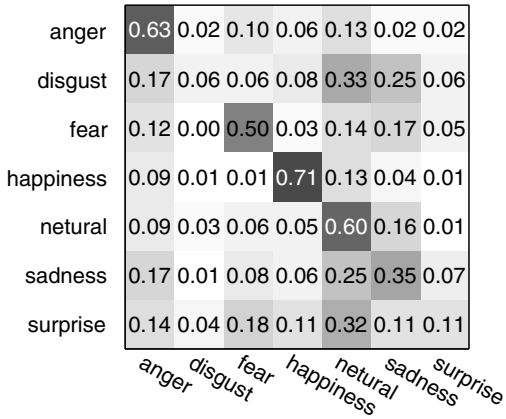


Fig. 5. The confusion matrix of the PCA fusion method based on Alex Net-2 and Gabor on the testing subset.

given on Fig.2 and Fig.3. The confusion matrix of the SKRRR fusion method and the PCA fusion method based on Alex Net-2 and Gabor on the testing subset are respectively given on Fig.4 and Fig.5.

In this paper, as we have not utilize some deep learning based methods to the audio, so the final recognition rate is not very satisfied, and we can improve the accuracy rate of our approach in future by utilizing some deep learning based methods such as DBN [42] to the audio.

V. Conclusions

In this paper, we present a bimodal emotion recognition approach using the convolutional neural networks and feature fusion method (the PCA fusion method, the KCFA fusion method and the SKRRR fusion method). The results on the testing subset of the AFEW 6.0 database show that the accuracy rate of the PCA fusion method and the SKRRR fusion method based on Alex Net-2 and Gabor are 53.46% and 50.93% with the svm classifier respectively, and are higher than the baseline of the EmotiW 2016 whose accuracy rate is 40.47%.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grants 61501249 and 41601601, the Natural Science Foundation of Jiangsu Province under Grant BK20150855, the Natural Science Foundation for Jiangsu Higher Education Institutions under Grant 15KJB510022, the Research Foundation of The Ministry of Education and China Mobile under Grant MCM20150504, The Open Foundation of Engineering Reasearch center of Widerand Wireless Communication Technology,Ministry of Education under Grant No. ZS002NY16002 and the NUPTSF under Grant NY214143.

References

[1] Y. Zong, W. Zheng, X. Huang, J. yan and T. Zhang. Transductive transfer LDA with riesz-based volume LBP for emoion recognition in the wild.

In Proceedings of the ACM International Conference on the Multimodal Interaction, pages 491-496, 2015.

[2] C. Shan, and R. Braspenning. Recognizing facial expressions automatically from video. *Handbook of Ambient Intelligence and Smart Environments*, pages 479-509, 2010.

[3] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39-58, 2009.

[4] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li and B. Wang. Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. *IEEE Transactions on Multimedia*, 18(7):1319-1329, 2016.

[5] Y. Tian, T. Kanade, and J.F. Cohn. Facial expression recognition. *Handbook of Face Recognition*, pages 487-519, 2011.

[6] A. Dhall, R. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 423-426, 2015.

[7] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 494-501, 2014.

[8] Z. Yu, and C. Zhang. Image based static facial expression recognition with multiple deep network learning. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 435-442, 2015.

[9] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):966-979, 2012.

[10] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. *In ACM International Workshop on Audio/visual emotion challenge*, pages 3-10, 2013.

[11] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie and M. Pantic. AV+EC 2015: the first affect recognition challenge bridging across audio, video, and physiological data. *In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3-8, 2015.

[12] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 461-466, 2014.

[13] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 509-516, 2013.

[14] B. Sun, L. Li, T. Zuo, et al. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 481-486, 2014.

[15] J.K. Chen, Z. Chen, Z. Chi, et al. Emotion recognition in the wild with feature fusion and multiple kernel learning. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 508-513, 2014.

[16] H.W. Ng, V.D. Nguyen, V. Vonikakis, et al. Deep learning for emotion recognition on small datasets using transfer learning. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 443-449, 2015.

[17] G. Levi, T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 503-510, 2015.

[18] A.C. Cruz. Quantification of Cinematography Semiotics for Video-based Facial Emotion Recognition in the EmotiW 2015 Grand Challenge. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 511-518, 2015.

[19] S.E. Kahou, V. Michalski, K. Konda, et al. Recurrent neural networks for emotion recognition in video. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 467-474, 2015.

[20] The Fouth Emotion Recognition in the Wild Challenge (EmotiW) 2016. https://sites.google.com/site/emotiw2016/challenge-details.

[21] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 19(3):00-34, 2012.

[22] A. Dhall, R. Goecke, J. Joshi and T. Gedeon. The Fouth Emotion Recognition in the Wild Challenge 2015: Baseline, data and protocol. *In Proceedings of the ACM International Conference on Multimodal Interaction*, 2016.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of IEEE*, 86(11):2278-2324, 1998.

[24] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1097-1105, 2012.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 1-42, 2015.

[26] S. Lawrence, C.L. Giles, A.C. Tsoi, et al. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98-113, 1997.

[27] A. Krizhevsky, G. Toderici, S. Shetty, et al. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725-1732, 2014.

[28] S. Ji, W. Xu, M. Yang, et al. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221-231, 2013.

[29] M.J, Lyons, J. Budynek, S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357-1362, 1999.

[30] W. Zheng, X. Zhou, C. Zou, et al. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Transactions on Neural Networks*, 17(1):233-238, 2006.

[31] F. Eyben, M. Wollmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *In Proceedings of the ACM International Conference on Multimedia*, pages 1459-1462, 2010.

[32] C. Shan, S. Gong, P.W. McOwan. Beyond facial expressions: learning human emotion from body gestures. *In Proceedings of the 2007 British Machine Vision Conference*, pages 1-10, 2007.

[33] Y. Wang, L. Guan, A.N. Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3):597-607, 2012.

[34] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. *In Computer VisionCACCV*, pages 143-157, 2014.

[35] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59-63, 2015.

[36] F. Eyben, M. Wollmer, and B. Schuller. OpenearaATintroducing the munich open-source emotion and affect recognition toolkit. *In Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 1-6, 2009.

[37] J. Yan, X. Wang, W. Gu, and L. Ma. Speech emotion recognition based on sparse representation. *Archives of Acoustics*, 38(4):465-470, 2013.

[38] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: affect recognition using multiple cues and modalities. *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185-218, 2008.

[39] J. Yan, W. Zheng, M. Xin, and J. Yan. Integrating facial expression and body gesture in videos for emotion recognition. *IEICE Transactions on Information and Systems*, E95-D(3):610-613, 2014.

[40] M. Liu, Z. Huang, et al. Partial least squares regression on grassmannian manifold for emotion recognition. *In Proceedings of the ACM International Conference on Multimodal Interaction*, pages 525-530, 2013.

[41] M. Turk, A.P. Pentland. Face recognition using eigenfaces. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586-591, 1991.

[42] M. Abdel-rahman, et al. Deep belief networks using discriminative features for phone recognition. *IEEE International Conference on Deep belief networks using discriminative features for phone recognition*, pages 5060-5063, 2011.

CPBL 教学法在本科教学中的实践分析:以交互设计课程为例

单美贤,谢皓薇,谢依蕾

(南京邮电大学 教育科学与技术学院,南京 210023)

【摘 要】 为了真正体现以学生为中心的教学理念,在面向本科生的"交互设计"课程教学中引入了基于项目的合作学习(CPBL)。文章从学生的自主学习能力、项目合作任务、团队合作过程等维度对大学生 CPBL 课程内容整体掌握程度的影响因素进行分析,结果表明学生的自主学习能力、团队合作过程对课程内容整体掌握程度有显著影响,而项目合作任务则无显著影响。

【关键词】 CPBL;自主学习能力;团队合作过程;项目合作任务;课程内容掌握程度

【中图分类号】 G642 【文章编号】 1003-8418(2017)03-0068-03

【文献标识码】 A 【DOI】 10.13236/j.cnki.jshe.2017.03.016

【作者简介】 单美贤(1975—),女,江苏张家港人,南京邮电大学教育科学与技术学院副教授、博士;谢皓薇(1992—),女,江苏南京人,南京邮电大学教育科学与技术学院硕士生;谢依蕾(1994—),女,江苏无锡人,南京邮电大学教育科学与技术学院硕士生。

一、研究背景

"基于项目的学习"最早是美国教育家克伯屈于 1918 年《设计教学法:在教育过程中有目的活动的应用》一文中提出,主张教学活动应以学生的需要为中心。基于项目的学习与传统教学模式的区别在于:以学生为中心,以一套独特且相互联系的项目(任务)为前提,借助多种资源,为实现项目目标为目的,强调过程性评价且评价主体多元化[1]。基于项目的学习与杜威的"做中学"以及布鲁纳的发现学习等思想有密切联系,同时它也在建构主义学习理论中找到了更扎实的理论依据。该教学方法发展至今,其研究成果非常丰富,实践应用案例也很多。近年来,我国教育工作者将其大范围应用于课堂教学实践中,大量研究结果从"管理层面、学习过程、能力培养和知识巩固"等视角证实了此学习方法的成效。

自主学习强调学生能够自己决定自己的学习目标,确定学习内容和进度,选择学习方法和技巧,监控学习过程及自我评价学习效果[2]。以学生自主学习为目标的"基于项目的学习"注重与生活以及社会实践相联系,可以是个体活动也可以是小组活动,大量的研究表明,小组合作学习有助于培养学生的团队合作、沟通和解决问题的能力。为了充分体现以学生为中心、更好地促进学生的自主学习,我们在"交互设计"课程教学中引入基于项目的合作学习(CPBL:Collaborative Project-Based Learning),以具体的项目为载体,使学生在实践中充分理解和体验交互设计的流程,注重过程性评价,以促进学生在团队合作中自主学习能力、团队合作与沟通能力、解决问题能力的提升。

二、CPBL 在"交互设计"课程中的具体实施

为了推动 CPBL 的有效进行,在"交互设计"课程的教学过程中,借鉴产品交互设计中的具体流程来组织教学内容。CPBL 具体实施步骤如下:

第一步,明确小组成员,确定项目任务。因为本课程面向的对象是大二本科生,他们之间相互了解,所以采取自愿的方式组成小组,小组人数为 3-6 人不等。明确小组成员之后,各小组采用"头脑风暴"的讨论方式,确定具体的交互设计项目主题,明确每个人在项目中的具体角色。

第二步,按照交互设计项目流程来推动项目的具体实施。

1. 发现需求阶段:为了确保项目定位准确并且具有一定的实践意义,在 CPBL 实施过程中,教师引导各小组成员从两个角度展开:一是用户角度:为了使交互设计产品真正满足用户需求,项目小组需要明确目标用户群体,并采取用户访谈、问卷调查等方法展开用户研究,挖掘用户需求,区分

什么是用户需求的表象、什么是用户需求的本质，形成用户需求文档。二是产品角度：主要采用竞品分析，即对竞争产品或者类似产品进行分析，形成产品体验设计文档，以找到改善自身产品的路径。

2. 分析需求阶段：在 CPBL 中指导各小组使用人物角色法，即根据用户研究获得的数据来生成关于用户描述性的模型——人物角色，最终形成目标用户模型，作为小组项目的工作目标对象。

3. 定义功能数据、进行信息架构和流程设计阶段：项目小组须理清设计任务，明确用户需要什么数据、用什么手段满足用户需求（即功能），进行信息架构设计，定义操作流程，全面呈现并分析用户与交互设计产品之间的关系，形成产品交互设计文档。

4. 原型设计：根据产品交互设计文档，项目小组以交互线框图的方式，讨论、验证、快速迭代改进，形成产品主要界面的原型。

5. 可用性测试：不同小组交叉对项目原型进行可用性测试，判断设计的产品原型是否与可用性原则相符，形成可用性测试报告。

第三步，成果展示和评价。每个小组展示项目实施过程中的过程性内容和最终作品，其他小组结合项目评价表给每个展示小组打分，分值占 40%，教师和助教的评分占 60%，构成项目组的团队成绩。个体成绩则以团队成绩（专业层面＋团队合作层面）为基础，适当结合小组成员的互评和个人自评的方式，这样即可把学生平时的合作能力、团队意识、学习态度纳入考核范围。

三、CPBL 实施情况的实证研究

（一）研究问题及研究方法

我们不难从“交互设计”课程的 CPBL 实施过程中发现，师生的角色发生了变化：学生是学习活动的主体，需要承担更多的责任来组织推动自己的学习，提高自主学习能力，在实际的项目情境中积极主动地参与小组合作任务，推动项目的顺利完成。教师不再是学习过程的控制者，而是项目实施的指导和促进者，在项目范围的界定、项目实施过程的推进及时间的分配等方面需要更多的参与，并平衡好课程教学目标与学生独特的项目解决方案。我们在基于“交互设计”课程 CPBL 实践的基础上，从学生的自主学习能力、项目合作任务（是否有明确的目标、清晰的流程等）、团队合作过程（分享、沟通、协调等）等维度对大学生 CPBL 课程内容整体掌握程度的影响因素进行实证分析，如图 1 所示。

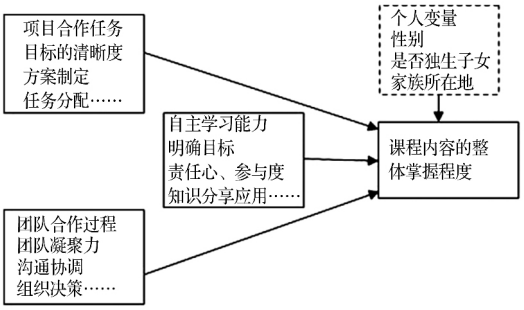


图 1　实证研究的结构

研究问题 1：学生的自主学习能力、项目合作任务、团队合作过程与课程内容整体掌握程度是否有显著的相关存在；

研究问题 2：个体差异（性别、独生子女与否、家庭所在地）的学习者，其自主学习能力、项目合作任务、团队合作过程与课程内容整体掌握程度是否有显著差异；

研究问题 3：学生性别、自主学习能力、项目合作任务、团队合作过程是否可以有效预测课程内容整体掌握程度，其预测力如何。

本研究在文献调研、行动研究、观察访谈的基础上，采用问卷调查的方式进行。问卷设计的依据是文章所提出的 3 个研究问题，共有 55 个项目（以李克特五点量表的形式测量）。主要研究对象是南京邮电大学某学院 2014 级本科生，发放问卷 192 份，其中有效问卷为 185 份，问卷有效率 96.4%。

（二）数据处理与分析

1. CPBL 中课程内容整体掌握程度的影响因素分析

量表采用的是 5 级评分表，从统计结果来看，CPBL 课程内容整体掌握程度平均值为 3.9288，接近于 4 分，说明 CPBL 有助于学生通过项目合作的形式较好地掌握课程内容。就自主学习能力来说，其每题得分平均值为 4.08763＞4；项目合作任务、团队合作过程的平均值分别为 3.47905、3.54375，居于中等程度范围。可见在 CPBL 中学生的自主学习能力得分较高，但项目合作过程中的方案制定、任务分配以及团队合作过程仍有待提高。

2. 学习者的个体差异对课程内容整体掌握程度的影响

采用独立样本 T 检验比较课程内容整体掌握程度总量表得分上是否存在性别、是否独生子

女、家庭所在地这三个方面的差异。检验结果表明：男生和女生在CPBL的课程内容整体掌握程度方面并无显著差异；独生子女在CPBL的课程内容整体掌握程度方面要优于非独生子女；学生的家庭所在地不影响其CPBL的课程内容整体掌握程度。

3. CPBL课程内容整体掌握程度影响路径

如图2所示，在对CPBL中课程内容整体掌握程度影响的路径中，有三条显著路径，一是自主学习能力→课程内容的整体掌握程度；二是团队合作过程→课程内容的整体掌握程度；三是自主学习能力→团队合作过程→课程内容的整体掌握程度。路径图的结果显示自主学习能力是对于CPBL课程内容整体掌握程度影响最重要的解释变量，不仅具有直接效果(8.137)，也具有间接效果(9.265×3.345)。

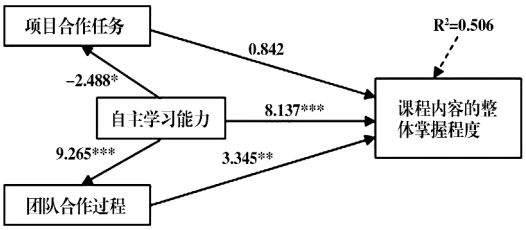


图2 路径分析图

另外，进行路径分析时，在模型中增加了控制变量即学生的个体差异(性别、是否独生子女、家庭所在地)，结果表明假设检验并没有因为这些控制变量的引入而变化，具体为年龄(β=0.002，p=0.975)、是否独生子女(β=-0.025，p=0.751)、家庭所在地(β=0.152，p=0.051)。

四、研究结论

本研究以本科教学“交互设计”课程中CPBL(基于项目的合作学习)的教学实践为依据，对CPBL课程内容整体掌握程度的影响因素进行了实证分析，包括学生的自主学习能力、项目合作任务、团队合作过程以及学生的个体差异性因素。

1. CPBL有助于学生通过项目合作的形式较好地掌握课程内容。较于传统的课堂教学而言，在基于项目的合作学习环境中学习者是项目的实施者和问题的解决者，能充分地发挥学习者的主动性和积极性。与此同时，CPBL教学实践过程中借助项目的形式把课程内容有机地整合，学习者不再是被动地掌握课程内容的知识点，而是在项目实施过程中应用这些知识点来解决问题，在掌握课程内容的同时提高了知识综合运用能力。

2. 自主学习能力对课程内容整体掌握程度有显著影响。在实证分析中，考察的自主学习能力主要包括：在小组合作中根据自己的实际情况制定学习目标和学习计划、根据合作需要进行有针对性的学习并认真完成自己的任务、解决合作中遇到的困难(寻求帮助、调整情绪、冲突应对)、自我评估及合作状态调整等方面的内容。这些因素对CPBL课程内容整体掌握程度有着显著的影响，因此在CPBL中培养学习者的自主学习能力，有助于提高其知识掌握与应用能力。

3. 项目合作任务对课程内容整体掌握程度没有显著影响。由于“交互设计”课程中的CPBL实施过程是按照交互设计项目流程来推动项目进展，各小组的目标清晰、任务的难易程度适中、制定的实施方案合理，因此项目合作任务对课程内容整体掌握程度没有显著影响。

4. 团队合作过程对课程内容整体掌握程度有显著影响。团队成员间的凝聚力、沟通协调能力、组织决策能力影响着团队合作过程，在项目的推进过程中，有些小组成员存在边缘化或不作为的现象，这与实证分析的结果是相一致的，即团队合作过程对课程内容整体掌握程度有显著影响。因此，提高团队成员间的凝聚力、培养团队成员的沟通协调和组织决策能力，有助于推进CPBL过程，进而较好地掌握课程内容。

5. 学习者的个体性差异对课程内容的整体掌握程度的影响。从实证分析中的数据可以看出，性别特征、家庭户口所在地对学习者的课程内容整体掌握程度无显著影响。学习者是否是独生子女对课程内容整体掌握程度有显著影响，独生子女的课程内容整体掌握程度优于非独生子女，统计数据表明，独生子女在自主学习能力和项目合作任务方面高于非独生子女，但其团队合作能力却低于非独生子女。

【参考文献】

[1]基于项目的学习简介[EB/OL]. [2016-7-17]. http://blog.sina.com.cn/s/blog_764c2e7d0101kaqb.html.

[2]孙晓玲. 课堂教学培养大学生自主学习能力的措施[J]. 江苏高教，2014(2)：94.

基金项目：2014年度江苏省社会科学基金项目(项目号：14SHD004)。

(责任编辑 朱旗)

Deep learning application: rubbish classification with aid of an android device

Sijiang Liu*[a], Bo Jiang[a], Jie Zhan[a]
[a]College of Educational Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

ABSTRACT

Deep learning is a very hot topic currently in pattern recognition and artificial intelligence researches. Aiming at the practical problem that people usually don't know correct classifications some rubbish should belong to, based on the powerful image classification ability of the deep learning method, we have designed a prototype system to help users to classify kinds of rubbish. Firstly the CaffeNet Model was adopted for our classification network training on the ImageNet dataset, and the trained network was deployed on a web server. Secondly an android app was developed for users to capture images of unclassified rubbish, upload images to the web server for analyzing backstage and retrieve the feedback, so that users can obtain the classification guide by an android device conveniently. Tests on our prototype system of rubbish classification show that: an image of one single type of rubbish with origin shape can be better used to judge its classification, while an image containing kinds of rubbish or rubbish with changed shape may fail to help users to decide rubbish's classification. However, the system still shows promising auxiliary function for rubbish classification if the network training strategy can be optimized further.

Keywords: Deep learning, image classification, mobile application, android development, rubbish classification

1. INTRODUCTION

Although the theory of multilayer neural networks has been widely studied in 1980s, it's been quiet for a long time until deep learning, the new machine learning pattern based on the former theory, was proposed in recent years. Due to the advantages of end to end learning, non-linear learning and good scalability[1], the deep learning method has made much progress in pattern recognition and artificial intelligence researches. Many classical problems benefit a lot from the deep learning idea, such as image classification[2-3], object detection[4-5], sense parsing[6], speech recognition[7], etc.

Meanwhile the progresses in theory also lead the advances in applications. The most famous event must be the program AlphaGo which defeated human players for the first time in Go chess. Researchers also develop deep learning applications to help blind people navigate unknown environments[8], or help users identify malware[9]. This paper focus on a similar practical problem in real life that people, confronted with kinds of rubbish, are usually like blind people, having difficulty in identifying rubbish's classifications and placing them in the right rubbish bins. Since the deep learning method has high accuracy rate in image classification and object detection areas, it can no doubt perform rubbish classification well if being integrated with other techniques by designing and implementing carefully.

There were some patents which described so called "smart bins" to help people classify rubbish automatically. The core techniques in those smart bins were traditional image process methods, and hardware composited by bloated modules needed to be installed in a rubbish bin. Obviously, this kind of design is out-of-date, inconvenient to maintain and easy to be damaged. Nowadays mainstream designs prefer to move much work to the cloud computing, leaving terminals light and handy. Therefore we hope that people can classify rubbish easier and faster with aid of an ordinary mobile device which almost everyone has. In this paper a prototype system using deep learning technique was proposed to realize the above goal.

The rest of this paper is organized as following. Section 2 gives an overview about our prototype system and the main workflow. Section 3 introduces some technique details. In section 4 some preliminary tests on the system are presented and discussed. Finally in the last section we conclude the paper and outline our future work.

*liusj@njupt.edu.cn;

Second International Workshop on Pattern Recognition, edited by Xudong Jiang, Masayuki Arai, Guojian Chen, Proc. of SPIE Vol. 10443, 104431P · © 2017 SPIE
CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2280285

2. SYSTEM DESIGN AND WORKFLOW

2.1 System architecture

As mentioned above, to realize the goal of rubbish classification conveniently, the prototype system in this paper should be designed with light and handy terminals, and much work should be done on "cloud", i.e., on a web server remote.

Considering that mobile devices have become important tools in our daily lives, it's natural to take advantage of them as terminals in our system. A simple and clean app on mobile devices could be enough if it has functions of taking rubbish pictures and uploading/downloading data stream to/from a server.

The remote web server will be responsible for requests from terminals, analyzing image data uploaded by them, and giving results back along its origin connection. In theory, the server should own a trained classification network, which uses an image containing rubbish as input and output an object detection result or a label of the rubbish type. On the basis of former output, the classification guide would be provided on users' mobile devices.

Hence the architecture of our prototype system corresponding to above descriptions is shown in figure 1.

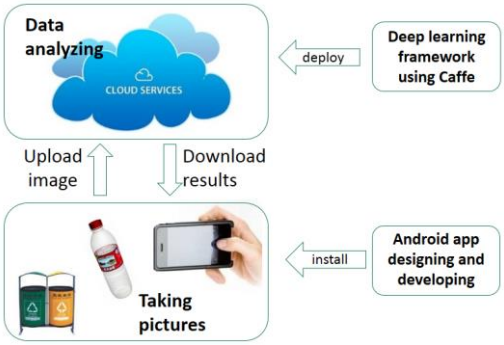


Figure 1. The architecture of the prototype system in this paper.

At least two significant advantages could be seen in this architecture. One is that cloud servers would collect massive image samples to persistently help improving classifying ability, and they are easily scalable with the number of terminals. The other is that light and handy terminals conform to the trend of internet of things, where terminals are desired to be with small size and light weight. The user terminals in this architecture need so simple functions that perhaps a small hardware module can perform the same work as well. So a rubbish bin integrated with that kind of module will be a more smart internet-of-things product.

2.2 Workflow of our android app

Since major work can be done by servers, terminals actually only have to do several quit simple jobs, including taking a picture of unclassified rubbish, uploading the image to servers for analyzing, retrieving the feedback and displaying results. We've developed an app organizing above jobs on android platform to guide users for smooth experience. It's not hard to extend this app to other mobile platforms.

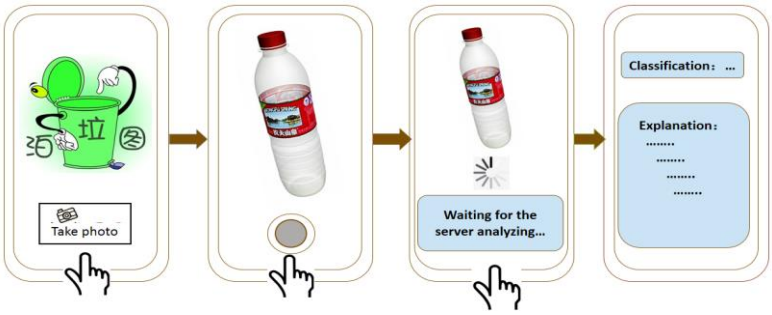


Figure 2. The workflow of the specially developed app on android platform.

3. TECHNIQUE DETAILS

3.1 Caffe framework for deep learning

"Caffe" is a clear and effective deep learning framework created by Dr. Yangqing Jia[10], and it's released under the BSD 2-Clause license. Many developers have forked Caffe source codes for improvements and researches.

A typical Caffe project is composed by two parts: network modeling and parameters configuration. The former defines the whole structure of the deep learning network as well as behaviors of every layer in it. The latter defines all parameters needed in network training such as weight decay, iteration number and so on. Both parts are implemented using Google Protocol Buffers[11] data format, a kind of textual modeling language, which hides implemental details and is platform-neutral. This feature is very friendly to developers.

A deep learning model, called Net in Caffe, should be combined by multiple layers to form a directed acyclic graph. Layers are basic units in a Caffe Net for computing and other data manipulations. One layer example is shown in figure 3 with its structure sketch on the left and Google Protocol Buffers description on the right.

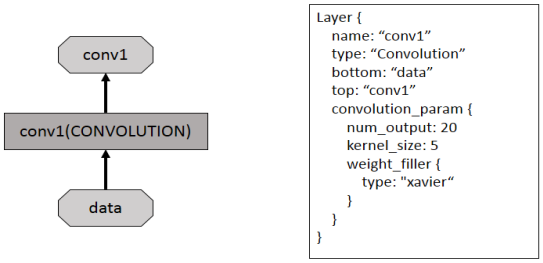


Figure 3. The structure sketch of an example layer (left) and its Google Protocol Buffers description (right).

Moreover, a Caffe Solver is indispensable to optimize the Caffe Net. That is to minimize a given loss function to obtain optimized parameters in the Caffe Net. Equation (1) shows a basic loss function, where *D* represents a dataset, and $X^{(i)}$ is an item in *D*. $f_W(X^{(i)})$ means the loss of an item. $r(W)$ is a regularization term to weaken the overfitting and λ is the weight. Stochastic gradient descent method can be used to update weights *W* set in the whole Caffe Net[12].

$$L(W) = \frac{1}{|D|}\sum_{i}^{|D|} f_W\left(X^{(i)}\right) + \lambda r(W) \tag{1}$$

Caffe Net training is an energy-consuming task, so the official website of Caffe has offered different kinds of trained Net[13]. This paper chose BVLC Reference CaffeNet trained on ImageNet dataset, and deployed it on a server following Caffe tutorials.

3.2 Post-classification method using regularization expressions

Theoretically, a Caffe Net for the goal of rubbish classification should be trained specially. Thus the server can offer classification services directly. However, since an already trained model was adopted mentioned above, the output of the Caffe Net corresponding to an input image would be class labels which cover most common objects in life, instead of actual rubbish classifications.

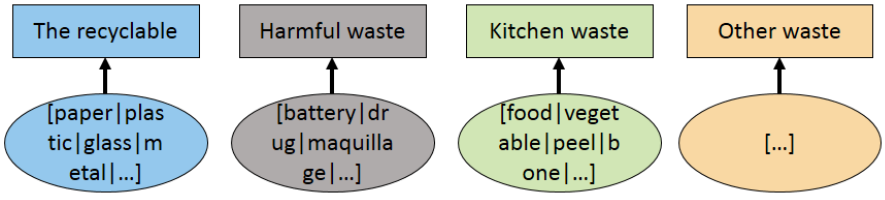


Figure 4. Four regularization expressions are associated with four rubbish classifications.

Therefore, the paper used regularization expressions as a post-classification method to match a class label to one specific rubbish classification. For example, there are four rubbish classifications in our system, shown in figure 4. And different regularization expressions including many labels are associated with those four classifications.

The output of the trained Caffe Net could be five class labels and their probabilities. In case the first probable label is a false prediction and make wrong classification, we use all five labels to help classifying and sum probabilities in four classifications respectively. The classification with the maximum sum value will be the final decision for an image input.

3.3 Android app developing

Since the official android developing IDE, Android Studio, was released in 2013, it has achieved rapid progress. More and more developers migrate their projects to Android Studio for its favorable usability and comprehensive supports.

In this paper we used Android Studio v2.2.3 and Android SDK 21 to develop our app, which was then installed on a Google Nexus 5 with android platform 5.0. The functions of this app are realized via four android activities and layouts. And it needs permissions to access the phone camera and network.

4. RESULTS ANALYSIS

We tested the system by three types of rubbish images: images containing a single object with its origin shape, images containing an object with changed shape and images containing several rubbish.

Obviously, tests using the first type of images can more probably get accurate classification because of the outstanding image classifying ability of the Caffe Net. One example is shown in figure 5. The top two class labels returned by the server were “vessel” and “container”, so our post-classification method judged the object in the image as the recyclable waste.

Figure 5. A test example using an image containing a glass.

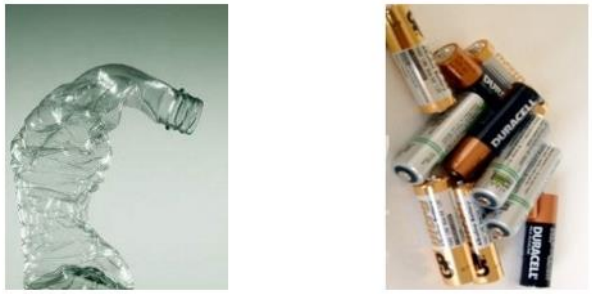


Figure 6. Test examples containing a damaged plastic bottle (left) and some batteries (right).

However, images containing an object with changed shape or containing several rubbish would cause wrong classifications, as shown in figure 6. On the left of figure 6 is a picture of a damaged plastic bottle. But two most probable labels returned by the server were “nematode” and “worm”, and the bottle was classified as other waste, which

should be the recyclable. While on the right of figure 6 is a picture containing some batteries, which should be harmful waste. But top two probable labels returned were “implement” and “device”, and rubbish in this image was finally classified as the recyclable incorrectly.

However, tests still show some feasibilities in our system. If users just take a photo of an object, the system can usually give a correct rubbish classification.

5. CONCLUSION

In this paper, we present a prototype system for the rubbish classification based on the deep learning method. Firstly the CaffeNet Model was adopted for our classification network and the trained network was deployed on a web server. Then an android app was developed for users to capture images of unclassified rubbish, interact with the server backstage and obtain the classification guide by a mobile device conveniently. Tests on the system showed its promising usability and directions for improvement. One critical factor is that the system adopted an already trained Caffe Net on the ImageNet dataset, so the classifying process is not performed directly. A post-classification step may cause unexpected errors although it can utilize the classifying ability of existing deep learning networks.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable suggestions to improve this paper. This work was sponsored by NUPTSF (Grant No.NY214199, NY214200 and NY215123).

REFERENCES

[1] Guo, X. X., Li, C., Mei, Q. Z., “Deep learning applied to games,” ACTA AUTOMATIC SINICA, 42(5), 676-684 (2016).
[2] Krizhevsky, A., Sutskever, I., Hinton, G. E., “ImageNet classification with deep convolutional neural networks,” Advances in neural information processing systems, 1097-1105 (2012).
[3] Simonyan, K., Zisserman, A., “Very deep convolutional networks for large-scale image recognition,” Proceedings of the 2014 International Conference on Learning Representations, (2014).
[4] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., “Overfeat: Integrated recognition, localization and detection using convolutional networks,” Proceedings of the 2013 International Conference on Learning Representations, (2013).
[5] Szegedy, C., Toshev, A., Erhan, D., “Deep neural networks for object detection,” Proceedings of the 2013 Advances in Neural Information Processing Systems, 2553-2561 (2013).
[6] Farabet, C., Couprie, C., Najman, L., LeCun, Y., “Learning hierarchical features for scene labeling,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1915-1929 (2013).
[7] Amodei, D., Anubhai, R., Battenberg, E., et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” arXiv preprint arXiv:1512.02595, (2015).
[8] Aljasem, D. K., Heeney, M., Gritti, A. P., et al., “On-the-Fly Image Classification to Help Blind People,” Intelligent Environments (IE), 2016 12th International Conference on. IEEE, 155-158 (2016).
[9] Kolosnjaji, B., Zarras, A., Webster, G., et al., “Deep Learning for Classification of Malware System Call Sequences,” Australasian Joint Conference on Artificial Intelligence. Springer International Publishing, 137-149 (2016).
[10] Jia, Y., Shelhamer, E., Donahue, J., et al., “Caffe: Convolutional architecture for fast feature embedding,” Proceedings of the 22nd ACM international conference on Multimedia. ACM, 675-678 (2014).
[11] Google Developers, “Protocol Buffers,” https://developers.google.com/protocol-buffers/, (2017).
[12] BVLC, “Solver,” http://caffe.berkeleyvision.org/tutorial/solver.html, (2017).
[13] BVLC, “Caffe Model Zoo,” http://caffe.berkeleyvision.org/model_zoo.html, (2017).

Enhancement of Color Image Based on Tone-Preserving

Ya'nan Yang[1], Xiaofan Wang[1], Feng Liu[1,2], Zongliang Gan[1,2]
1. Jiangsu Province Key Lab on Image Processing & Image Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
Corresponding author: liuf@njupt.edu.cn

***Abstract*—Since the lighting conditions in strong contrast regions between the light and dark cant be estimated accurately by traditional center/surround Retinex algorithm, the over-enhancement and color distortion may exist. In view of this, combining with the human visual characteristics, a color image enhancement algorithm based on tone-preserving was proposed. A determination function was added to the bilateral filter to estimate illuminance image more accurately and weaken over-enhancement. According to human visual masking effect, the improved gamma correction was utilized to correct the brightness of illumination image adaptively and the local contrast of reflection image obtained by division was enhanced based on local statistics. Besides, the final enhanced image was obtained by combining illumination image with reflection image, which can make image appear more natural. Compared with other similar algorithms from both subjective and objective aspects, the results show that this method being applied to low-contrast color image enhancement can not only improve image clarity, but reduce color distortion.**

***Keywords*—Retinex; tone-preserving; bilateral filter; masking effect; local contrast**

I. Introduction

In the actual lighting conditions,the images acquired by digital cameras, camcorders, smart phones or other terminals always appear to be too bright or dark in certain regions, resulting in low overall contrasts and unsatisfactory visual effects. The traditional image enhancement methods such as histogram equalization, homomorphic filtering and nonlinear mapping are mostly applied to gray images. Color image enhancement is even more complex with respect to the gray image. As the color image contains color information, the color should also keep undistorted when the image brightness and details are improved. Furthermore, enhancing RGB three components respectively with the same methods will damage correlation between them and easily lead to color distortion. Currently, space conversion is the more commonly used method. That is, the image is converted to a particular color space so as to separate brightness from color, and its brightness is only processed to maintain hue consistency. However, which method of enhancing image brightness to pick is still a subject worthy of study.

At present, among a variety of color image enhancement algorithms, the Retinex (abbreviation of retina "Retina" and the cerebral cortex "Cortex") algorithm [1] has been more widely used because it can achieve a good balance between the dynamic range compression, edge enhancement and color constancy. Retinex algorithm includes random path Retinex, MsCann's Retinex, variational Retinex and center/surround Retinex. Among them, center/surround Retinex is currently the most widely used algorithm, including SSR (Single Scale Retinex), MSR (Multi-scale Retinex) and MSRCR (Multi-Scale Retinex with Color Restoration). However, as the three algorithms all assume uniform incident light change, the over-enhancement and color distortion may be easily produced in light mutation regions. To solve these problem, reference [2] used bilateral filter rather than the traditional Gaussian filter to reduce over-enhancement and have more accurate illumination estimation; reference [3] applied MSR to the luminance component of HSI color space and keep S unchanged, in order to keep color consistent; for reducing color distortion, reference [4] not only applied MSR to the luminance component of HSV color space, but corrected the saturation component S, aiming to maintain the relative relationship between brightness and saturation.

In this paper, a new method based on tone-preserving was proposed. That is, the image is converted to HSI color space, only its brightness is enhanced while its color component H and S keep unchanged. The experiments show that this method can effectively reduce over-enhancement and color distortion.

II. Retinex Theory

According to Retinex theory [1], an image can be represented as:

$$S(x,y) = R(x,y) \times L(x,y) \tag{1}$$

Where *L(x,y)* is illumination image, *R(x,y)* is reflection image corresponding to image's essential attribute, which can be obtained by making the equation (1) logarithmic, and subtracting luminance image from the original image *S(x,y)*.

$$log(S(x,y)) = log(R(x,y) \times L(x,y)) \tag{2}$$

$$log(R(x,y)) = log(S(x,y)) - log(L(x,y))) \tag{3}$$

The basic idea of center/surround Retinex is convoluting the original image with a Gaussian kernel function to obtain illuminance image. For example, SSR algorithm [5, 6] uses a

978-1-5090-2860-3/16/$31.00 ©2016 IEEE

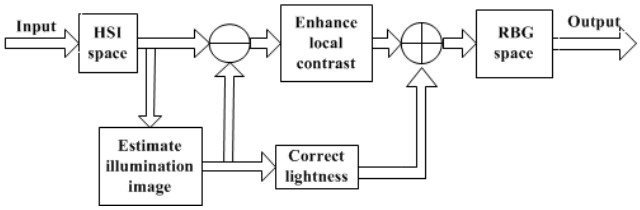


Fig. 1: Algorithm block diagram

single-scale Gaussian kernel function to estimate illuminance image.

$$R_{SSR}(x,y)) = log(S(x,y)) - log(S(x,y) \otimes F(x,y)) \quad (4)$$

Where $\otimes$ represents convolution, *F(x,y)* is Gaussian kernel function. Due to the problem that single scale parameter is difficult to balance color fidelity and contrast enhancement, Rahman proposed MSR algorithm [7, 8], as shown in formula (5):

$$R_{MSR,i} = \sum_{k=1}^{K} W_k \times R_{SSR,i} \quad (5)$$

Where W_k is weighting coefficient, *k* is number of Gaussian function or surround scale. For color images, $R_{SSR,i}$ is the SSR result of scale *k* and channel *i*.

The MSRCR algorithm[9, 10] uses a set of color recovery factors and multiplies by the result of MSR as in formula (6) and (7), which can make up for deficiencies in color fidelity to a certain extent.

$$R_{MSRCR,i}(x,y) = G_i(x,y) \times R_{MSR,i}(x,y) \quad (6)$$

$$G_i(x,y) = \beta\{log(\alpha \times S_i(x,y)) - log[\sum_{i=1}^{N} S_i(x,y)]\} \quad (7)$$

Where G_i(x,y) is color recovery factor, α is the intensity of nonlinear transformation control. β is gain constant. A lot of experiments demonstrate that, although the color recovery factor can increase color saturation, there is still some color distortion in light mutation regions.

III. The Proposed Algorithm

In order to keep tone constant, we convert image to HSI color space based on human visual characteristics. First, for the reduction of over-enhancement, we extract illuminance image with improved bilateral filter and obtain reflection image by division operation; next, we correct the luminance image brightness using improved gamma function and enhance the contrasts of reflection image with local statistics; finally, combining illumination image with reflection image, we convert the enhanced image back to RGB color space. The algorithm block diagram is shown in Fig.1:

A. Illumination Estimation

Gaussian kernel function only considering the location of surrounding pixels, thus easily leads to inaccurate illumination estimation in light mutation regions. On the other hand, the bilateral filter [11] considering both the position and corresponding values of the pixels, can estimate illumination more accurately. However, when applied to regions under strong contrasting light cases, the estimated illumination value is always smaller than its real value and the reflected value is bigger, which will cause the over-enhancement.

Therefore, a determination function is added to bilateral filter to determine whether a pixel can participate in filtering or not. More specifically, if a surrounding pixel value is greater than or equal to the center pixel value, the value of determination function is 1, otherwise, the function value is 0. In this way, it can ensure the estimated illuminance value is equal to or larger than its true value and thus weaken the over-enhancement. The improved bilateral filter is described as in formula (8):

$$L(p) = \frac{\sum_{q\in\Omega} S(p)c(p,q)g(S(p),S(q))t(S(p),S(q))}{\sum_{q\in\Omega} c(p,q)g(S(p),S(q))t(S(p),S(q))} \quad (8)$$

$$t(S(p),S(q)) = \begin{cases} 1, & \text{S(p)}\geq \text{S(q)} \\ 0, & \text{else} \end{cases}$$

$$c(p,q) = e^{-\frac{1}{2}(\frac{d(p,q)}{\delta_d})^2}$$

$$g(S(p),S(q)) = e^{-\frac{1}{2}(\frac{\delta(S(p),S(q))}{\delta_\gamma})^2}$$

Where *t(S(p),S(q))* is determination function, *g(S(p),S(q))* and *c(p,q)* are respectively brightness and distance similarity between surrounding point *p* and center point *q*; *d(p,q)* is Euclidean distance; $\delta(S(p),S(q))$ is brightness difference; *L(p)* is the value of point p after being filtered; *p* is the location of the point whose coordinates is (x, y); *q* is the position of neighborhood pixels of point *p*; Ω is filter child window. In addition, the original image *S(x,y)* needs to be normalized to a range of 0 and 1, so as to correct illumination image brightness.

B. Brightness Correction

Illumination image mainly contains light information corresponding to the low frequency parts, and the increment of brightness can highlight details in dark regions. Furthermore, according to human visual masking effect [12], the human eyes are more sensitive to texture details under medium brightness background than those under the extremely high or low brightness cases. The traditional gamma correction can be used to improve image brightness globally, and the larger the correction factor is, the more obvious the effects leading to over-enhancement of the relatively brighter regions are. Therefore, combining with the initial brightness value of luminance image, the gamma correction factor is revised to correct brightness adaptively as shown in formula (9). From Fig.2, the improved gamma correction can be used to achieve more substantial enhancement of low light regions, the lower

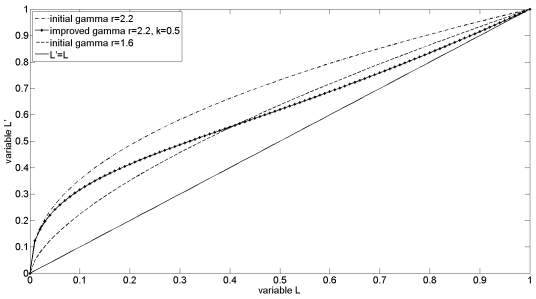


Fig. 2: Gamma correction curve

amplitude increase of bright regions and almost constant of the highlighted regions, so as to avoid suppressing lighter region details while highlighting the dark region details.

$$L'(x,y) = L(x,y)^{\frac{1}{r}+ksin(L(x,y))} \tag{9}$$

Where *L'(x,y)* is illumination image after brightness correction, *r* is gamma correction factor whose value is normally greater than 1, *k* is scale factor ranging from 0 to 1.

C. Local contrast enhancement

The logarithmic transformation is used to compress dynamic range by classical Retinex algorithm, although the overall contrast is enhanced, the local contrast is reduced.As a result, we directly utilize division operation to obtain reflection image. As reflection image represents details and color information, the local contrast enhancement of reflection image can improve image clarity and visual effect. The local contrast enhancement based on local statistics [13] is described as follows:

$$R(x,y) = \frac{S(x,y)}{L(x,y)} \tag{10}$$

$$R'(x,y) = \overline{R}(x,y) + k \times [R(x,y) - \overline{R}(x,y)] \tag{11}$$

Where R'(x,y) is output reflection image, $\overline{R}$ is neighborhood average of 3×3 region, *k* is gain factor between 1 and 2, the larger *k* is, the more obvious the enhancement effects are. Local variance or local contrast can be improved k^2 times, which can be deduced from formula (11).

D. Combination of Illumination and Reflection Image

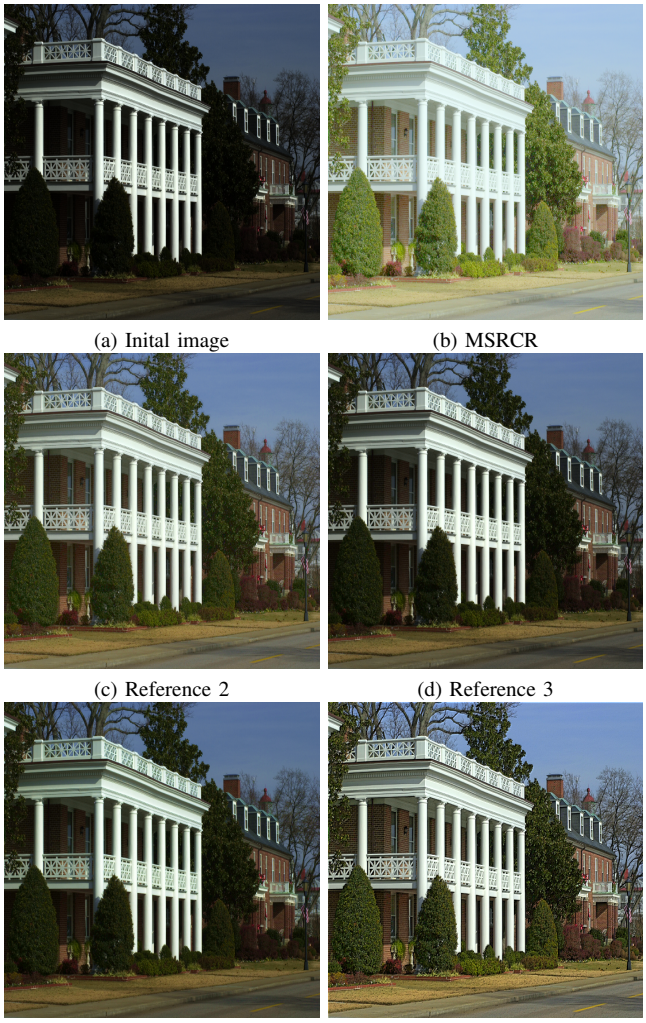
Generally, the illumination image also contains a portion of high-frequency information, if being completely removed, the results may inevitably lose some details and the color will appear unnatural. For this reason, we obtain the enhanced result by combining the luminance image in formula (9) with reflection image in formula (11) to obtain the enhanced result.

$$R_{last}(x,y) = R'(x,y) \times L'(x,y) \tag{12}$$

Where R_{last}*(x,y)* is the enhanced compound image.

IV. Analysis of Experiment

To validate the effectiveness of the proposed algorithm, we choose three relatively low-contrast images for comparative analysis. Compared with MSRCR, reference [2], [3], [4], the results are shown from Fig.3 to Fig.5.



(a) Inital image (b) MSRCR

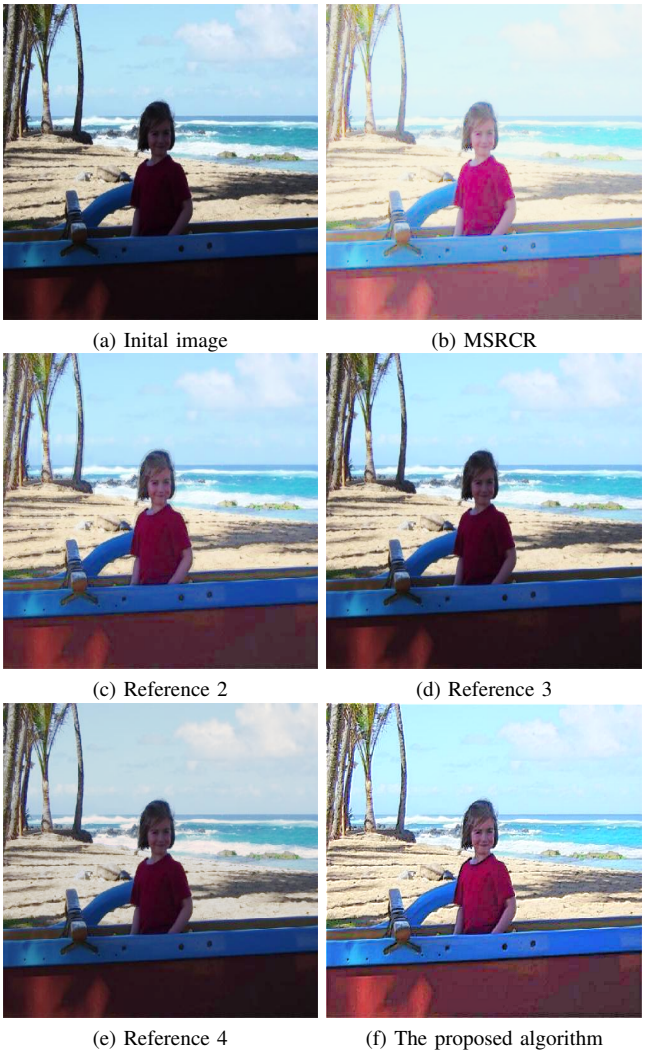
(c) Reference 2 (d) Reference 3

(e) Reference 4 (f) The proposed algorithm

Fig. 3: Comparison 1 of results

A. Subjective Analysis

As can be seen from Fig.3 (b) to Fig.5 (b), the dark details and contrasts are enhanced by MSRCR to a certain extent. However, the over-enhancement results in the reductions of local contrast especially in Fig.3 (b). Besides, there is color distortion in Fig.4 (b) and Fig.5 (b). Comparing Fig.(d) with Fig.(e) from Fig.3 to Fig.5, the subjective enhancement effects are similar in preference [3] and preference [4], although the dark region details are enhanced, it is not so obvious for the darker region in Fig.3(d) and Fig.3(e). From Fig.3 (c) to Fig.5 (c), the visual effects in preference [2] are superior to preference [3] and preference [4] regardless of the brightness or details. But the halo still occurs in strong shading regions as shown in Fig.4(c). From Fig.3(f) to Fig.4(f), the proposed algorithm not only improves the overall brightness and details of dark regions and suppresses over-enhancement effectively, but also makes the image color more vivid and overall visual

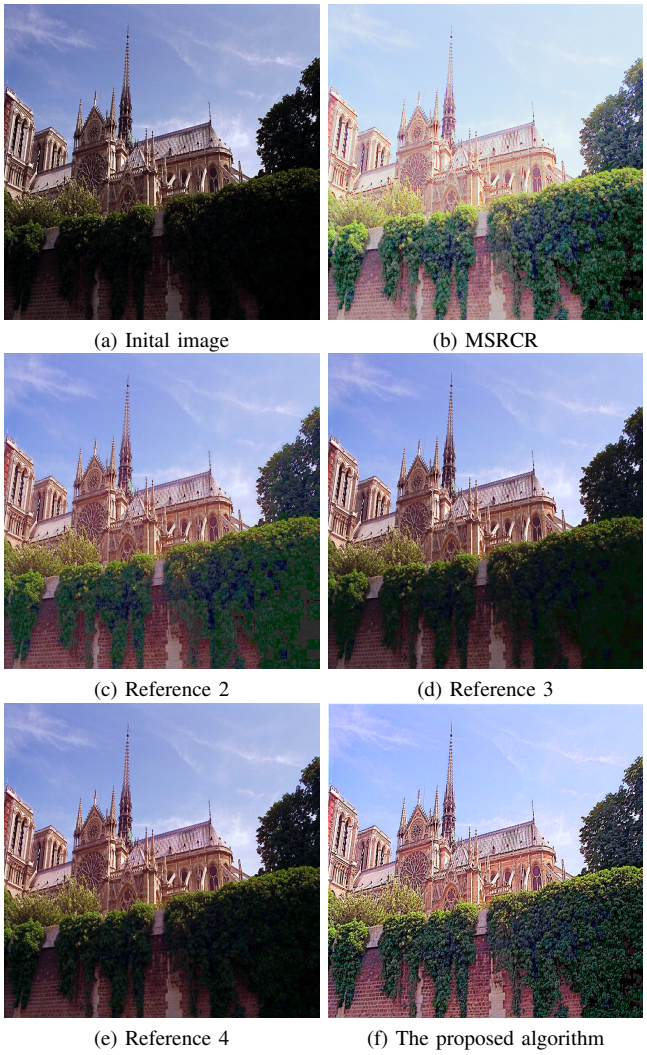


(a) Inital image (b) MSRCR

(c) Reference 2 (d) Reference 3

(e) Reference 4 (f) The proposed algorithm

Fig. 4: Comparison 2 of results



(a) Inital image (b) MSRCR

(c) Reference 2 (d) Reference 3

(e) Reference 4 (f) The proposed algorithm

Fig. 5: Comparison 3 of results

effect better. It is superior to the preference [2] in the aspects of brightness, sharpness and color.

B. Objective Analysis

In order to further demonstrate enhancement effects, we choose information entropy, mean and average gradient as objective evaluation indexes. The entropy evaluates the amount of information contained in the image. Mean reflects the image average brightness. Average gradient reflects the image details.

As can be seen from Table I, the information entropy, mean and average gradient of all algorithms are improved. Among them, MSRCR owes highest mean because of over-enhancement, but the information entropy and average gradient are similar to other algorithms except the proposed algorithm. There is little difference between preference [3] and preference [4], and the three indicators of preference [2] are higher than theirs. The information entropy and mean of the proposed algorithm are higher than preference [2], and the average gradient is significantly higher, which means the overall brightness is higher and the detail enhancement is more obvious.

V. Conclusion

The main contribution of this paper is the improvement of bilateral filter extracting illumination image more accurately, which can weaken over-enhancement. According to human visual masking effect, the brightness of illumination image is corrected adaptively and the local contrasts of reflection image are also enhanced to further improve clarity. The combination of illumination image and reflection image can make the result more real and natural. The experiments show the proposed algorithm not only reduces the over-enhancement, but maintains the consistency of tone, and the overall visual effect is much better than other algorithms.

Acknowledgment

This work is supported by National Natural Science Foundation of China (NSFC) (61501260, 61471201), Natural Science Foundation of Jiangsu Province (BK20130867), Jiang-

TABLE I: OBJECTIVE EVALUATION INDEXES

Algorithms	Fig.3			Fig.4			Fig.5		
	Entropy	**Mean**	**Gradient**	**Entropy**	**Mean**	**Gradient**	**Entropy**	**Mean**	**Gradient**
Original image	7.0	57	8.5	6.5	56	3.8	6.8	56	3.5
MSRCR	7.4	156	10.7	6.6	172	4.7	6.9	164	3.9
Reference[2]	7.4	100	11.0	6.9	105	5.4	7.2	106	4.9
Reference[3]	7.2	69	10.1	6.9	73	4.8	7.0	69	4.3
Reference[4]	7.2	66	10.0	6.8	68	4.7	7.0	65	4.4
This paper	7.7	105	19.3	7.4	120	9.0	7.6	136	8.4

su Province Higher Education Institutions Natural Science Research Key Grant Project (13KJA510004), The peak of six talents in Jiangsu Province (2014-DZXX-008), Natural Science Foundation of NUPT (NY214031), and "1311 Talent Program" of NUPT.

REFERENCES

[1] E. H. Land and J. J. Mccann, "Lightness and retinex theory," *Journal of the Optical Society of America*, vol. 61, no. 1, pp. 1–11, 1971.

[2] C. Chao, "Improved single scale retinex algorithm in image enhancement," *Journal of Computer Applications and Software*, vol. 30, no. 4, pp. 55–57, 2013.

[3] J. Yan and K. Zhang, "Hsi space based on illumination information in multi-scale retinex image enhancement algorithm," *Computer Engineering & Application*, vol. 46, no. 23, pp. 31–33, 2010.

[4] X. Qin, H. Wang, Y. Du, H. Zhen, and L. Zhenhua, "Structured light image enhancement algorithms based on retinex in hsv color space," *Journal of Computer Aided Design & Computer Graphics*, vol. 25, no. 4, pp. 488–493, 2013.

[5] K. Yao and D. Tian, "Shadow removal from images using an improved single-scale retinex color restoration algorithm," in *International Joint Conference on Computational Sciences & Optimization*, 2009, pp. 934–938.

[6] Z. Al-Ameen and G. Sulong, "A new algorithm for improving the low contrast of computed tomography images using tuned brightness controlled single-scale retinex," *Scanning*, vol. 37, no. 2, pp. 116–125, 2015.

[7] Z. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *International Conference on Image Processing*, vol. 3, 1996, pp. 1003–1006.

[8] C. H. Lee, J. L. Shih, C. C. Lien, and C. C. Han, "Adaptive multiscale retinex for image contrast enhancement," in *International Conference on Signal-Image Technology & Internet-Based Systems*, 2013, pp. 43–50.

[9] D. J. Jobson, Z. U. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *Image Processing IEEE Transactions on*, vol. 6, no. 7, pp. 965–976, 1997.

[10] S. Zhang, P. Zeng, X. Luo, and H. Zhen, "Multi-scale retinex with color restoration and detail compensation," *Journal of Xi'an Jiaotong University*, vol. 46, no. 4, pp. 32–37, 2012.

[11] Y. Wang, C. Yin, Y. Huang, and H. Wang, "Defogging image based on bilateral filtering," *Journal of Image and Graphics*, vol. 19, no. 3, 2014.

[12] Y. Zhang, D. Agrafiotis, M. Naccari, M. Mrak, and D. R. Bull, "Visual masking phenomena with high dynamic range content," in *IEEE International Conference on Image Processing*, 2013, pp. 2284–2288.

[13] G. Liu and X. Meng, "A new local statistical active contour model and algorithm based on sar image," *Geomatics and Information Science of Wuhan University*, vol. 40, no. 5, pp. 628–631, 2015.

Online Video Synopsis Method through Simple Tube Projection Strategy

Jing Jin[1], Feng Liu[1,2,*], Zongliang Gan[1,2], Ziguan Cui[1,2]
1. Jiangsu Province Key Lab on Image Processing & Image Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
liuf@njupt.edu.cn

***Abstract*—Traditional video synopsis methods model the processing into an optimization formula where relations among objects such as collision cost are utilized while entire re-calculation is introduced under each possible temporal shift. Unlike the pairwise cost optimization, we propose a low-complexity and efficient online synopsis method where each tube is processed independently. Without tubes' comparison, the rearrangement is accomplished by a simple projection strategy and an updating projection matrix which records the newest information of the moving space. Furthermore, buffer and a predefined fitness condition also help to increase spatial and temporal utilization. Experiment results demonstrate that the proposed method is superior to other synopsis methods in the processing speed and temporal consistency.**

***Keywords*—video synopsis; projection matrix; buffer; collision**

I. Introduction

In the world armed with cameras that generate massive records at an explosive speed 24 hours constantly, the volumes of surveillance videos are burgeoning. Surveillance videos are usually too lengthy to get fully utilized. It is also extremely time consuming and ineffective to browse and search by human force for interesting objects.

Many approaches are proposed to condense the volume of videos and relieve the memory burden at the same time. Among them, video synopsis is a convenient and user friendly method that grabs widely attentions at this field. It aims to provide a compact video representation, while preserving the essential activities of the original video [1].

There are basically two approaches of video synopsis processing, including offline and online method. Offline video synopsis is firstly proposed by Pritch et al [1, 2] and developed as a two-phase processing method [3, 4]. The first phase is to scan the whole video in advance for capturing and storing trajectories and background. On the second phase, all tubes are rearranged wholly at one time by formed unary and pairwise cost functions. Consequently, the offline framework has the drawbacks of huge calculation and huge memory cost when it deals with long videos. All these factors hinder this way from practical usage. For the online synopsis method, it means that the preparing stage and optimization procedure are processed parallelly. By introducing buffers to store temporal results that are unable to be processed in time, the paper [5, 6] transforms the global optimization as an approximated stepwise optimization. As a result, it significantly reduce temporal complexity while re-calculation is still needed under each possible temporal shift. Another paper [7] takes a new insight to condense videos in an fully online manner. By maximizing a posterior estimation composed by three models, objects' instances are formed into full trajectories. Beside, a synopsis table is introduced to reallocate each tubes without complicated cost optimization. The processing speed is much fast and the collision among moving objects are largely avoided. However, this framework sometimes also cause unsatisfactory situations due to the unpredictable nature of instances during tubes' trajectories.

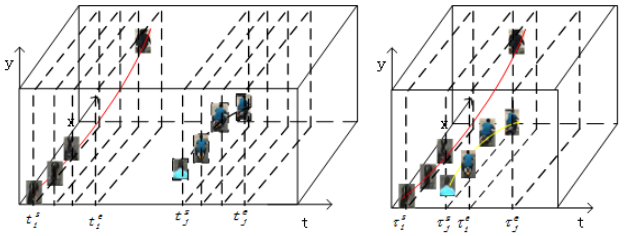
In this paper, we propose a low-complexity and efficient online synopsis method to compress videos online. In our framework, tube is the basic processing unit. After the online detection and tracking, the vanished tube is rearranged by a simple tube projection strategy and a projection matrix. Recording the newest information of the moving space, the updating projection matrix helps to avoid collision and costs' re-calculation. Consequently, original videos can be condensed into synopsis videos with little collision and temporal disorder very quickly. Buffer and a predefined fitness condition are also introduced to generate more compact results. The proposed method is compared with other two synopsis methods among several test videos from three metrics of $speed$, $chronological\ disorder$ and $condensation\ ratio$. Experimental results validates that our framework can condense input videos with lower complexity and higher efficiency.

II. Proposed method

A. Overview of Proposed Framework

The aim of video synopsis is to generate compact representation for original videos, as shown in Fig. 1. In Fig. 1(a), the cube stands for moving space of an video. Red and yellow curves represent respectively a walking girl with black coat from t_i^s to t_i^e and a riding man with blue suit from t_j^s to t_j^e. These two person show up sequentially with a relatively large time interval which cause much redundancy. While after objects are rearranged into new temporal positions and useless frames are removed, the synopsis video shown in Fig.1 (b) is much more condensed.

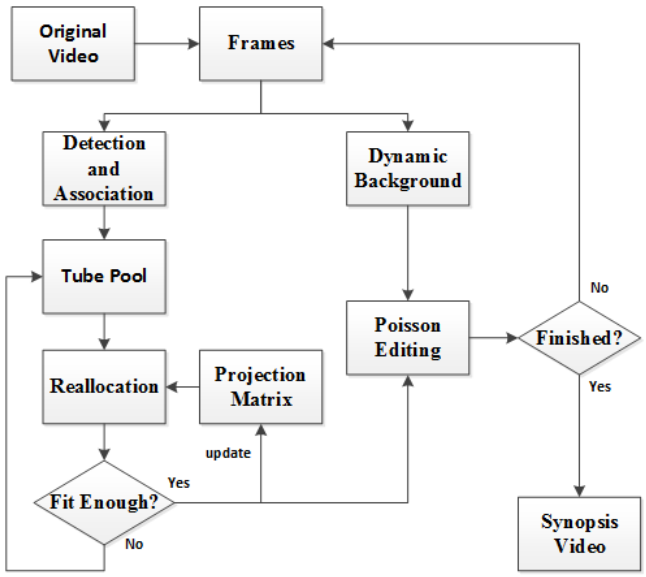
978-1-5090-2860-3/16/$31.00 ©2016 IEEE



(a) Original Video (b) Synopsis Video

Fig. 1: Illustration of video synopsis

The framework of our method is illustrated in Fig. 2. All coming frames of a given original video are sequentially detected and associated to form full trajectories of moving objects. The associated results are stored in a tube pool. Meanwhile, dynamic background images by averaging frames during a certain time interval are generated for later stitching. Once a tube vanishes in the moving scene, it can be rearranged to an ideal position by a projection matrix which records the newest information during reallocation. Once the new position satisfies the predefined fitness condition, the projection matrix is then updated and the tube is stitched into the dynamic background by Poisson editing technology [8]. Otherwise it should be added into the tube pool for next processing. The final synopsis video is then generated by the iteratively performed of this procedure.



(a)

Fig. 2: Framework of Proposed Method

B. Dynamic Tube Generation

Tube is the processing unit in this paper and is composed by all instances during the trajectory. Tube i is represented as $Tube_i = \{O_i^s, O_i^{s+1}, ..., O_i^e\}$, where O_i^s stands for the instance of object i appears at frame t. Each instance is a rectangle mask and indicates the location. $I(x, y, t)$ represents the pixel value of original video at and location (x, y) and frame t, satisfying $(1 \leq x \leq W, 1 \leq y \leq H, 1 \leq t \leq T)$. Objects' detection and tracking should be performed from beginning until last. Considering the robustness and effectiveness, we apply fast R-CNN [9] to detect all instances at each frame. Meanwhile a fused distance which is used for association into full trajectories is defined as follows:

$$D(O_i^{t_1}, O_i^{t_2}) = d_{spatial}(O_i^{t_1}, O_i^{t_2}) * d_{chisqr}(H_i^{t_1}, H_i^{t_2}) \quad (1)$$

where $d_{spatial}(O_i^{t_1}, O_i^{t_2})$ is spatial distance of two instances and $d_{chisqr}(H_i^{t_1}, H_i^{t_2})$ is Chi-square distance of two histograms in the content. With a simply greedy algorithm, the instance at current frame is connected to the nearest instance that appears before. It should be noticed that fast R-CNN could not enable full detection at one hundred percent ratio and will also generate a little broken tubes. However, it still can extract enough important information to represent original videos. In addition, the broken situation is also considered in the progress of distances calculation and is optimized at the code level.

Besides the online tracking and association, changing background for the synopsis video should also be recorded along time. In this paper, we only deal with surveillance videos with cameras that are standstill, where the difference in background is largely due to the change of illumination. Consequently, the average pixel values during fixed time interval of input video is assumed as the background image for objects stitching [10].

C. Simple Projection Strategy

The main idea of video synopsis is to remove activity-less frames, and rearrange tubes in video frames to make objects that appear sequentially in an original video can appear simultaneously in the shortened video [6]. While the high condensation rate will lead to severe collision artifacts and inconsistency of temporal sequence. They are the two influential items that determine the quality of synopsis videos and viewers' watching feeling.

For two shifted objects and each relative time shift between them, collision cost is usually defined as the volume of their space-time overlap among the shared time period as follows:

$$E_c(i, j) = \sum_{t \in \tau_i \cap \tau_j} e^t(i, j) \quad (2)$$

$e^t(i, j)$ records the shared pixel number of object i and j at the shared frame t. While from another aspect, it can also be understood as the volume of collided pixels of two or more instances that appear at same frame in the rearranged video just as below:

$$E_c = \sum_{i,j} E_c(i, j) = \sum_{(x,y,\tau)} \xi(x, y, \tau) \quad (3)$$

$$\xi(x, y, \tau) = \begin{cases} 1 & (x, y) \in O_i^\tau, \ \sharp i \geq 2 \\ 0 & other \end{cases} \quad (4)$$

The above equation illustrates the collision cost from above two different aspects. Intuitively, as long as the instances that

share common pixels don't show up at same frame, then collision among them wont exist, This trick means that the shared pixels should belong to different objects at different frames to avoid collision.

As a consequence, the rearrangement of a moving object should satisfies the above condition at all pixels of each instances during the trajectories. Based on above consideration, a mapping matrix M with the same size as video's two dimensional moving space is introduced. It records the newest temporal positions of each pixels in the moving space. Consequently in order to avoid collisions, the reallocated temporal position τ_i^s of object i should ensure all sequential instances won't occupy any pixels at the latest recorded temporal positions according to matrix M.

$$\tau_i^s = \max_{t\in\{1,\ldots,t_i^e-t_i^s+1\}} \max_{(i,j)\in O_i^t} (M(i,j)-t) \tag{5}$$

After the above reallocation, new positions during object i's trajectory will be occupied. So the projection matrix should be updated accordingly as follows and t has the same range as Equation 5:

$$M(i,j) = \begin{cases} \tau_i^s + t & (i,j) \in O_i^t \\ M(i,j) & other \end{cases} \tag{6}$$

Intuitively, the projection matrix M can be understood as continuous mapping of original videos during the condensation process. And tubes' reallocation is accomplished by the trajectories' information and the already projected data of previous processed frames. This is what we infer the simple projection strategy here.

Temporal consistency cost creates a preference for maintaining the temporal relations between objects by penalizing cases where relations are violated. Even though we process basically according to the vanish sequence of objects, as long as there are rooms for incoming tubes, the showing sequence still could be alternated. Consequently, after each reallocation, increasing the minimum value of matrix M will ensure the following objects will never show up before already processed objects. However, this mechanism will lower the ratio of spatial usage generally.

D. Optimization

As mentioned before, the maintain of temporal consistency will lower the spatial and temporal utilization ratio. So our method make a balance among them with buffer and a predefined fitness condition. During the reallocation, the changing of reallocation environment will leads to different fitness values for a given object.

To evaluate the quality of a tube to be stitched into synopsis video under projection matrix, we also form a matrix W with the same size as matrix M as follows:

$$W(i,j) = \frac{\max(M) - M(i,j)}{\max(M) - \min(M) + 1} \tag{7}$$

As shown in the above equation, matrix W represents the suitability of each pixels among all possible temporal position. For tube i, it's fitness can be easily transformed as the maximum value in matrix W among all pixels during the trajectory:

$$fit(tube_i) = \max_{t\in\{1,2,\ldots,L_i\}} W(O_i^t) \tag{8}$$

Consequently, the value of $fit(tube_i)$ and a predefined threshold which is set as 0.5 in our experiment would determine the suitability of tube i. Tube i can be allocated and stitched if the fitness value is higher than the predefined threshold. Otherwise, it will be added in the buffer until it is suitable enough to be processed. Such mechanism can exploit the resources of each position in temporal and spatial space to the best extent. Although it distract the temporal continuity more or less, it is especially suit for generating compact and short videos with less redundancy.

III. Results and Discussion

A. Experiment Setting and Results

In order to evaluate the performance of proposed method, we have tested several surveillance videos on an Intel Core i5 computer with a 3.2 GHz CPU and 8GB RAM. As for the space in this paper is limited, so we can't describe all the tested videos fully in detail. Consequently four representative videos among them are selected, whose detailed information is listed in the Table I. The first frames of the selected four videos are also showed in the Fig. 3 to give a concrete describe, By the way, we should notice that not only temporal duration but also resolution and values of frames per second (FPS) of Video 4 are larger than any other three test videos, which could absolutely need more calculation time and spaces during the condensation.

(a) Video1 (b) Video2

(c) Video3 (d) Video4

Fig. 3: The first frame of four selected videos

The sampled images of selected original videos and corresponding synopsis videos are shown in Fig. 4. At each column, the top three pictures are from original video and many spaces

(a) Frames about Video1 (b) Frames about Video2 (c) Frames about Video3 (d) Frames about Video4

Fig. 4: Frames from four selected original and synopsis videos

TABLE I: Detailed information of four selected videos

Videos' Name	Resolution	Duration	FPS	♯Frames
Video 1	320×240	00:05:10	18	5592
Video 2	320×240	00:05:00	18	3755
Video 3	320×240	00:05:37	18	6081
Video 4	1920×1080	00:41:30	25	52084

are empty among them, while the moving space of synopsis video becomes more occupied and condensed shown below.

Three metrics are adopted for quantitative evaluation: *speed*, *chronological disorder* and *condensation ratio*. We also implement offline method [3] and high performance stepwise optimization method [6] as comparative methods. Detailed experimental results are shown in the Table II and Table III.

TABLE II: Comparisons of speed and chronological disorder

Videos' Name	Method [3]		Method [6]		Proposed Method	
	Speed	CD	Speed	CD	Speed	CD
Video 1	8.67	10.00	247	3.22	568	1.44
Video 2	8.15	11.25	143	5.38	337	2.00
Video 3	8.53	10.39	129	8.64	310	2.63
Video 4	0.35	25.48	7.61	28.20	15.6	6.22

Speed refers to the average processed number of frames per second in average. Higher speed implies faster processing and low-complexity which are crucial for practical usage. As Table II shows, our method can deal with the largest number of frames per second, while offline framework has the lowest processing speed. We can also see that the speed is decreased with the increased video content and pixel resolution. Thats the reason why the last test video is processed relatively slower than the other three videos. It should also be noticed that as the framework [3] is a two-phase optimization, so the time spent during detection and tracking is not considered in speed value calculation for all three methods.

Chronological disorder (CD) is a metric to evaluate the ability of maintain temporal relations between objects. Synopsis videos with lower CD values ensure a better sequence display and would not disturb viewers' understanding for original videos that heavily. Like but different form temporal consistency cost which is defined in [11], CD is defined as the ratio between the number of disordered object pairs and all moving objects' number:

$$CD = \frac{1}{\sharp Tube} \sum_{i,j \in Tube} \{1 | D^o(l_i^s, l_j^s) * D^s(l_i^\tau, l_j^\tau)\} \quad (9)$$

where $D^o(l_i^s, l_j^s)$ and $D^s(l_i^\tau, l_j^\tau)$ are respectively the frame distance of tube i and tube j that showing in the original and synopsis video at first time. $\sharp Tube$ refers to the number of all moving objects. Tubes are rearranged basically according to their vanish sequence in our method, which maintain a good content sequence of original video. As Table II shows, our method has lower CD values respective to Method [6]. Besides Method [3] has the worst temporal consistency because objects' shift can take values from the whole temporal range in this framework.

TABLE III: Comparisons of condensation ratio

Videos' Name	Method [6]		Proposed Method	
	♯Frames	CR	♯Frames	CR
Video 1	736	7.59	1749	3.197
Video 2	958	3.92	1417	2.65
Video 3	1031	5.9	3406	1.785
Video 4	6124	8.5	17856	2.922

Condensation ratio (CR) is defined as ratio between the number of frames in original and synopsis videos. Higher CR values implies more condensed videos. In the offline framework, as the length of synopsis video should be manually assigned before optimization, it is set the same value as our method for comparison of other metrics. As a result, the data of metric CD about Method [3] is not listed here. As Table III shows, stepwise synopsis method has higher CR values and condensed results than our method. However, we

should also mention that the higher condensation performance of Method [6] is achieved by the expense of lower processing speed and larger chronological disorder than our method.

B. Discussion

Above experimental results validate the efficiency of proposed method in the metrics of speed and chronological disorder.

As for the computational complexity, reallocation stage in the framework of proposed method plays an important role to increase the overall performance. Each tube is rearranged by a simple projection strategy and don't need complex energy optimization among objects. As a result, the computational time for a certain tube is linear with the sum of areas occupied during it's trajectory and the size of projection matrix. However, traditional video synopsis methods formulate the rearrangement into a pairwise cost formula. During the cost optimization, re-calculation should be conducted under each possible temporal shift between the two compared tubes. So the computational complexity for each tube under such framework is dependent on the product values of range of temporal shift and sum of areas occupied by the selected two tubes. As for the temporal consistency, our method deal with moving objects basically according to their vanish sequence, which can maintain the better temporal information to some extent than other methods. As a whole, it is our rearrangement strategy that leads to the improved performance.

IV. CONCLUSION

A low-complexity and efficient online synopsis method has been proposed in this paper. Our framework differs much with other pairwise cost optimization method. A projection matrix is introduced to record the newest temporal information of moving space and avoid collisions between tubes. Along with a simple projection strategy, each moving object can be rearranged into new temporal position extremely fast without complicated comparison with other objects. Besides, buffer mechanism and a predefined fitness condition are also applied to increase the spatial and temporal utilization. The advantages in the low-complexity and temporal consistency ensure the framework a practical solution to condense lengthy surveillance videos.

V. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (NSFC) (61501260, 61471201), Natural Science Foundation of Jiangsu Province (BK20130867), Jiangsu Province Higher Education Institutions Natural Science Research Key Grant Project (13KJA510004), The peak of six talents in Jiangsu Province (2014-DZXX-008), Natural Science Foundation of NUPT (NY214031), and "1311 Talent Program" of NUPT.

REFERENCES

[1] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 435–441.

[2] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[3] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[4] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered synopsis of surveillance video," in *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, 2009, pp. 195–200.

[5] S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Online content-aware video condensation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2082–2087.

[6] J. Zhu, S. Feng, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "High-performance video condensation system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 7, pp. 1113–1124, 2015.

[7] C.-R. Huang, P.-C. J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, "Maximum a posteriori probability estimation for online surveillance video synopsis," *IEEE Transactions on circuits and systems for video technology*, vol. 24, no. 8, pp. 1417–1429, 2014.

[8] P. Prez, M. Gangnet, and A. Blake, "Poisson image editing," *Acm Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[9] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[10] C.-R. Huang, H.-C. Chen, and P.-C. Chung, "Online surveillance video synopsis," in *2012 IEEE International Symposium on Circuits and Systems*. IEEE, 2012, pp. 1843–1846.

[11] W. Fu, J. Wang, L. Gui, H. Lu, and S. Ma, "Online video synopsis of structured motion," *Neurocomputing*, vol. 135, pp. 155–162, 2014.

公共空间新媒体艺术观念与表现研究①

陈媛媛（南京邮电大学 传媒与艺术学院，江苏 南京 210046）

[摘 要]公共空间新媒体艺术的观念内涵呈现多向度范式，表现为比特空间的延伸，社会认知的转向以及艺术互动隐喻的象征；其叙事及互动体验的参与体现了虚拟与现实空间的多维交叉，感性的逻辑建构以及肢体语言的表达参与。新媒体艺术的进入，形成公共空间新媒体艺术的互动接口，以达到艺术表现、沟通和互动的目的。

[关键词]公共空间新媒体艺术；新媒体艺术观念；新媒体艺术表现

[中图分类号] J9　**[文献标识码]**A　**[文章编号]**1008-9675（2016）06-0176-05

一、公共空间新媒体艺术观念内涵的多向度范式

公共空间新媒体艺术是一种复合艺术形式，是存在于公共空间的艺术与新媒体艺术的互动美学内涵发生联系的一种艺术表现形式，是公共空间的艺术与当下社会公众产生沟通的一种方式，以新媒体技术为表现手段；从美学角度来说是通过新媒体艺术的互动形式体现“互动”“自由自在”的美学内涵，从社会学意义来说是一个社会的人文化和社会化的形象代表，规范“人机共栖”的社会形态。作为其表现形态的观念内涵，表现出全新的交流和构建范式。

（一）公共空间新媒体艺术的观念内涵

相较于广义的新媒体艺术形式，公共空间新媒体艺术的观念内涵、美学思想仍可追溯至科学的发明与大量运用；不仅改变人类的物质生活，也改变着人类的精神生活。纵观百年工业革命的科技史，每一次科技进步，都震撼人们的思维传统变化。每一次科技进步都为艺术提供“新视野”、“新观念”和“新思维”发展的契机。正如摄影术催生了“技术美学”；电影和动画开拓了“世界的艺术”；机械和动力启发了“装置艺术”；电视和摄像则孕育“录像艺术”等。每当新技术的产生或成熟，就会产生各种新艺术流派和各种“新美学”。19 世纪至今，艺术运动的不断产生，新媒体艺术的发展，几乎就是一部现代科技史。

从西方艺术史看，艺术往往与科学的革命联系在一起。设计常被人们视作科学与技术的“结晶”，也被看做是连接生产与审美的“桥梁”。现代新媒体艺术的滋生土壤与工业和信息产品的“美学”设计需求相关。从历史文化角度做探讨，19 世纪末 20 世界初，科学与工业技术的发展，使绘画、雕塑等传统艺术发生功能性变化，形成现代艺术、影像艺术和设计艺术三大独立的艺术状态，三个艺术形态相互渗透，相互影响，形成交融和互动的新艺术。在西方古典美学解构的同时，后现代主义思想逐渐形成，并不断掀起对传统美学观念的冲击，在 20 世纪 60 年代达到高潮。由此，新媒体艺术形成以“生活审美化”和“艺术大众化”为核心的新的艺术形式。

新媒体艺术的美学思想，主要来源于蒙德里安、康定斯基的抽象主义艺术、包豪斯的工业设计思想、意大利马里内蒂“未来主义”的“机器美学”、法国杜尚达达主义（dada, 木马意思）的对欧洲传统文化厌弃。除此之外，19 世纪英国的拉斯金和莫里斯发起“新手工艺美术运动”的设计风格，主张艺术要与日常生活结合，艺术家应该从工作室走入生活，创作大众所理解和喜爱的作品；“生活即艺术”、“艺术和反理性艺术”思想，20 世纪 50—60 年代，动力艺术和光效应艺术、20 世纪 70 年代的录像装置技术，数字合成技术等，都对新媒体艺术美学思想造成影响。20 世纪 50 年代，各种艺术运动将传统艺术推向各种新型城市空间，艺术家的偶然性、机遇及参与性等举动加强了城市的活化性，形成环境的媒体化，如奥托 . 皮纳（Otto Piene）的“天空艺术”，克里斯托和詹妮 . 克

收稿日期：2016-08-20

作者简介：陈媛媛（1986— ），女，江苏南京人，南京邮电大学传媒与艺术学院讲师，研究方向：新媒体艺术理论与创作研究、动画艺术创作与研究。

①基金项目：2015 年度教育部人文社会科学研究青年基金《公共空间新媒体艺术的构成及应用研究》（15YJC760014）；2014 年度南京邮电大学人文社会科学研究基金项目《大数据时代新媒体艺术互动形态社会应用研究》（NYS214016）。

劳德夫妇(Christo and Jeanne Claude)的“大地艺术”等。

此外，英国与美国发展出杜尚为代表新的艺术流派“波普艺术”，提出“通俗的、短暂的、可消费的、低廉的、大批量生产的、年轻的、妙趣诙谐的、性感的、诡秘狡诈的、有魅力的、大生意的”，和现代媒体艺术的表现和思维方式存在密切逻辑关系；其更贴近日常生活利用大众传媒，依靠复制技术创作，挑战传统的艺术定义的审美思想。

综上，公共空间新媒体艺术作为一种新的艺术形式，给人们提供了广阔的艺术空间；艺术空间的拓展，又使当代审美的外延与内涵得到明显的扩大和延伸。应用领域的社会信息服务业、工业设计标准信息服务、智能娱乐产品和高级数字娱乐产品设计开发等，已使社会形成“数字化”的认同。伴随社会“数字化”的发展，文化从最初狭窄的特定艺术种类，扩张到人类所有的精神领域和意识领域，日常生活开始趋于审美化。

（二）观念内涵的多向度范式表现

1. 比特空间的延伸

在社会的日常生活中，空间是容积、能够容纳体积的代名词，它是和实体的物理相对存在的，人们对于空间的感受是借助实体触摸而得到的。在社会及人类的发展中，人们会用围合或分割或圈占来取得自己所得的空间；与此同时，空间的封闭和开放是相对的。公共空间新媒体艺术以比特信息为基础、以各种不同的形式塑造空间，改变空间，从而使人产生不同的感受。

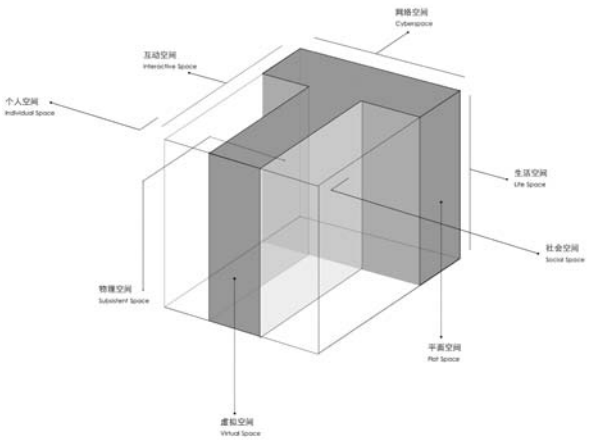


图1 公共空间新媒体艺术的空间多维形式构成

公共空间新媒体艺术所塑造的空间存在着垂直与水平维度；在上图的维度模型里，这些被系统化的空间塑造存在互有交叉，有共同的部分，也有不同的部分，对于其内涵与范畴有不同的定义和延伸，并形成了不同的空间领域；但此领域可被打破和重叠。

（1）平面空间向多维立体空间扩展

相较于新媒体艺术，传统的艺术视觉表达通常是平面设计元素为主。而公共空间新媒体艺术则更多的是将平面设计元素进行无限延伸，塑造视觉通透和渐层互动呈现，以此来延展平面艺术设计的纵向空间。公共空间新媒体艺术多是将平面艺术空间以更直观的方式转换、穿插，将有限的、静态的空间集合通过装置性立体艺术拓展成无限的、动态的空间元素集合，营造出立体的贴近真实的感受。

（2）新型的公共空间形成，物理空间与虚拟空间交叉并置。

公共空间新媒体艺术以其独特的实体数码展示方式，充分运用观众与艺术的互动，使观众沉浸其间，浑然不觉真实与虚幻之分；观众通过对此新媒体艺术的感受，游走于充满节奏的多元数码世界，感受独特的艺术视觉效果，沉浸于抽象和无限相互交织的时空。这种相互交织的时间与空间概念，即带来新型的公共空间形成，处于物理空间与虚拟空间交叉并置之中。

（3）信息交流是空间的主体，个人空间和社会空间混合。

在空间方面来说，公共空间新媒体艺术所塑造的空间，其主体是信息交流，其交流的对象是个人或群体的精神文化，所以其空间拓展也表现为个人空间和社会空间的混合。公共空间新媒体艺术的表现形式，大多是在特定的时空环境中，将人类日常生活中已消费或未消费过的物质文化实体，进行艺术的有效选择、利用、改造、组合及演绎，以展示出新的个体或群体丰富的精神文化意蕴。此空间必须是能使观众置身其中的、具体“环境”特性的空间。公共空间新媒体艺术在空间范式的概念延伸，成为观众个体信息与社会群体信息互相交流的环境，成为可传递动态讯息的环境构造体。

2. 社会认知的转向

在电子邮件和网络电话的世界里，我们共同生活的传播空间无穷大，其强度和速度是几年前难以想象的。新媒体艺术不再是特权阶层的精神盛宴，其艺术主张已为民众普遍认同。“网络创意群”、“以人为本”的虚拟社会，人类追求新鲜刺激图像信息的兴致及个人表演欲得需求也产生了价值。“复制拼贴”，“异类合成”，“再创意”“恶搞”等社会现象，由此引发更多的文化争论、价值观变化，以及社会认知的转向。

一方面，公共空间新媒体艺术艺术与非艺术的边界被破除，艺术的领域变宽，艺术的创作者和接受者不再以典型化式、静态的沉思替代艺术结果，艺术的阅读方式和图像的判断能力不再借助经典为蓝本，而是面对当下，直接从自下而上真实的情景中去接受，去体验，甚至是参与其中，互为表里。从杜尚“波普艺术”的“艺术即生活”观始，艺术已从画廊、沙龙、博物馆等四壁围困的模式走向户外，走向空间化；其表现形态从平面走向空间，从单一的画种走向绘画媒介的混合性过渡，更多表现为静态与动态的光效应、

肢体和观念等非传统的艺术出现，尤其是公共空间新媒体艺术已将视觉图式的共性特征被个性化的文化索求所解构。

另一方面，公共空间新媒体艺术的传播，表现手法和表现风格不断在改变，也改变着大众的审美情趣和审美要求。“虚拟”视觉与“实境”感觉同步，是大众在真实社会与虚拟情景交替环境下不断追求的刺激视觉体验；这种新体验与新认知，在人体感官作用下，将大众由被动的接受者，转向为可以按照个性化的审美定位，通过各种手段对艺术方式进行改变和再创造。个人也可以是艺术家，艺术设计与技术的艺术融合，具有更强的渗透力和感染力。艺术家不再像传统艺术那样关起门来独立地完成一件作品，而是更多地以多人合作的方式进行创作，这一点特别体现在艺术家与不同领域的技术人员的合作方面。传统意义中作为观看者的观众地位的改变。观众不再是被动的观看者，而成为了引发作品开始运作的关键因素，或直接成为作品的一部分。

3. 隐喻的象征

艺术主要的职责之一，不是再现世界，而是通过对世界的再现，使我们以特定的态度和特殊的角度去看这个世界。……所谓产生效果，也即让那些投身在世界中的观众对世界的看法产生转变，或进一步得到肯定。公共空间新媒体艺术是艺术，也具有文化的隐喻象征意义；有着赖以产生的社会基础，归根到底是社会化分工的产物。

公共空间新媒体艺术是形象化较强的文化产业。商业性不是其第一和唯一属性，但其艺术的文化灵魂是其文化价值成功的关键所在。公共空间新媒体艺术的艺术象征意义，是一个由物质到精神的生产转换过程。作为在新媒体中的艺术形式，不仅是观众艺术消费的过程，还始终伴随着观众的消费与再创作过程。公共空间新媒体艺术拥有艺术的独特魅力，不仅具有深远的艺术价值，衍生出来的文化含义蕴藏着比作品本身更大的文化价值和商业价值，具有特定的文化隐喻象征意义。其文化隐喻象征意义基础及表现在于：

1、具有科学技术商品化载体的特性。公共空间新媒体艺术是大众文化、大众艺术，对新媒体艺术价值的接受与认同。因此，这种形态的艺术应当符合观众不同群体，多元化的视觉经验和自然规律，使大众可以轻松地体验到艺术表现之下所要表达的感情和意义。

2、以娱乐、消遣游戏、精神抚慰为目的。以公共空间新媒体艺术特有介于真实和虚幻之间独特的美学特点，虚构内容，巧思创意，表现与现实落差的体验空间或创意性情境，达到“游戏冲动”的精神体验。“游戏冲动”在席勒的美学思想中，是一个重要的概念。……游戏的根本特点，就在于自由活动。席勒正是这样来理解游戏的，他说：“我们说一个人游戏，是说他审美地观照自然，并创作了艺术，把自然对象都看成是生气灌注的。在这里面，单纯的自然的必然性，让位给了各种能力的自由活动；精神自发地与自然相和谐，形式与物质相和谐。”（第十五信）[1]。公共空间新媒体艺术通过技术传递手法，从文化涵义，年龄结构等方面，带给大众视觉上创意想象，带来娱乐性和幽默诙谐、轻松明快，刺激和好奇，愉悦的美感和享受，以此实现文化隐喻中寻求人和自然和谐相处方式的价值需要。

二、公共空间新媒体艺术身体叙事及互动体验的参与

（一）空间性的多维性：虚拟与现实

在新媒体技术实现“虚拟假定”的空间世界里，公共空间新媒体艺术表现手法已不受制于现实规则；其艺术的形式更具视觉冲击力，如何直接、具体、个性化地创造出美术视觉语言，技术地传达设计理念和文化、艺术主张，实现技术与艺术完美结合，已成为设计师创作始终不渝地寻找、挖掘的追求。公共空间新媒体艺术的空间多维形式构成的基础在于：

1、空间的形式延伸及新思维的挖掘。新媒体技术带给社会的变革不仅仅是“一切”，更多的是人脑的延伸和扩展，尤其是无法替代设计的创造性思维。在公共空间新媒体艺术具象表现中，需要技术，运用技术，但在理解技术的同时，需要设计思维和启发创意灵感，掌握日新月异的计算机硬件和软件技术，并灵活把握视觉新语汇的表达，运用计算机技术系统，支撑美术设计；包括对点、线、面、方、圆、三角等几何形和各种色彩的元素，根据设计指令进行平移、重复、渐变、旋转、错位、变异、增减等操作，生成理想的图形和色彩；构成各种最基本设计元素的储存，用艺术与技术的“语言”表达思想，真正达到“人机对话”的艺术境界。

2、“艺术”和“技术”的双重关系。首先，“艺术”是“技术”实现的主导与核心。公共空间新媒体艺术的理念与创意，源于创作者主观文化艺术修养，既包含创作者对客观世界的认识，表现了对美和丑的审视和审美价值观念的表达；其创意与实现，也存在艺术思想与和技术“鼠标”“键盘”等形式下的“交互对话”的操作关系，同样具备人脑与电脑操纵“人机对话”的特点。在公共空间新媒体艺术形态展现中，人脑艺术表现完全处于主宰和支配的地位，技术是辅助人脑更有效地发挥其主观能动性和智慧的工具，使其迅速地在显示屏上设计，并通过多次修整达到理想的图形，便捷地把设计的创意传达给程序。其次，“技术”是实现“艺术”主张的辅助工具。技术的进步与发展，为公共空间新媒体艺术的艺术形态、创作、审美、设计手段及最终的视觉传播效果，带来崭新的思维空间。

其设计创意，既有艺术思维，也有技术的思维；而关键技术体现在计算机动画制作硬件与软件运用上。不同的公共空间新媒体艺术设计效果，取决于不同的硬件、软件功能。虽然制作的复杂程度不同，但其基本原理是一致的。

（二）互动系统与模式建构：感性的逻辑建构

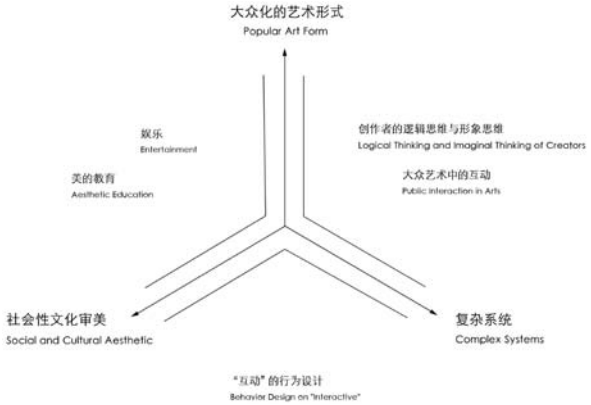


图2 公共空间新媒体艺术感性的逻辑建构

其逻辑建构主要表现为以下三方面，并互为因果，相辅相成：

1、公共空间新媒体艺术是一种大众化艺术形式，即包含艺术创作者的逻辑思维与形象思维、与大众在艺术中的“互动”。思维不管是形象思维或逻辑思维，都是认识的一种深化，是人的认识的理性阶段达到的对事物的本质的把握。形象思维的过程在实质上与逻辑思维相同，也是从现象到本质，从感性到理性的认识过程。但这过程又有与逻辑思维不同的、本身独有的一些规律和特点，这就是在整个过程中思维永远离不开感性形象的活力和想象。在表现形式上，无论传统艺术还是计算机艺术虚拟的水彩、素描、油画、雕塑等艺术手段及任何材料，都可以成为公共空间新媒体艺术“人机互动”实现形式构成元素，成为创作者与大众发挥艺术审美灵感实现的途径。

2、复杂系统与人互动的方式成为可能。相对于传统媒介的单向信息传递，新媒体与多维度和非线性，信息的双向、乃至多维的传递、反馈、碰撞、融合和激发，即构成了“交互与互动”的内容。

大型互动装置艺术《云》(“Cloud”) 是由 6000 个新的或废旧的灯泡组成，于 2014 年在卡尔加里 (Calgary) 城市第一个白夜节做公开展示。作为创建该装置过程的一部分，创作者收集当地的家庭、企业、博物馆和生态站烧坏白炽灯泡。该互动装置艺术最初的想法是创建社区和艺术家之间的非正式合作，并以降低成本的方式做实验性艺术创作的后期使用试验。无数的灯泡组成了一朵巨大的云朵，每个灯泡都装上拉弦，让观众来控制该灯泡云的照明。观众通过即兴的拉线开关与作品互动，使自己成为装置艺术集体中

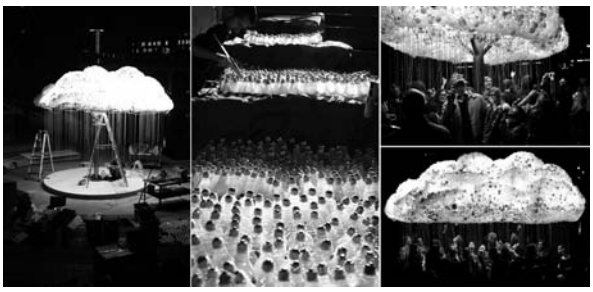


图3 大型互动装置艺术《云》（“Cloud”）装置现场，以及观众的互动

的个体进行交互。

《云》(“Cloud”) 互动装置艺术充分体现了复杂系统中人的“互动”行为设计。公共空间新媒体艺术“互动”关注的重点不是形式与内涵，而是“互动”的行为设计。传统设计的设计元素是点、线、面、色彩、肌理等，而公共空间新媒体艺术“互动”设计的元素是：复杂系统与人互动的方式——即人的行为。

3、公共空间新媒体艺术文化审美的社会性。在审美王国里，每个公民都是‘自由公民’，在权利力量的国度中，人和人以力相遇，他的活动受限制；在安于职守的、伦理的国度里，人和人以法律的威严相对峙，他的意志受到束缚；在审美的国度中，人就只需以形象显现给别人，只作为游戏的对象而与人相处，通过自由去给予自由，这就是审美王国的基本法律。美成了物质与精神、感性与理性、客观与主观之间的中介。美在人性发展的过程中，所起的正是这样的一种中介作用。正是这种中介作者用，使它能够改造人、教育人、克服人的片面性，使人成为具有完整的人性的人。

高度的娱乐性成为公共空间新媒体艺术的主要功能之一。在新媒体信息传播趋于图形化、动态化、互动化的现代社会形态中，公共空间新媒体艺术以其独特的文化形式与审美价值被更多的人认知。在“人机互动”在线形态下，时空特性淡化了各种文化时空观，差异文化交流将统一到同一虚拟时空里，公共空间的互动审美得到最生动的表现。基于互联网等新媒体平台多路径、多选择、多结局实现艺术交互和审美互动，包括各种有效的互动方式、增强和扩充，关注大众的心理和行为特点的普遍性、代入性以及互动操作过程参与感，都使大众在欣赏过程中充分发挥移情效应，最终大众的审美艺术观发生变迁。

（三）肢体语言的表达

相较于原手工传统技术技艺经验的艺术形态，从表现形式的结构逻辑以及相关审美趣味方面看，不断发展的公共空间新媒体艺术在新视觉中兴起的公共空间新媒体艺术，在手段上区别于以物质媒介为手段的视觉经验，更多地运用肢体语言处理媒介技术，所产生视觉经验已是截然不同的文化界面。

公共空间新媒体艺术中肢体语言的加入使得传统

图4 2010年8月25日，在德国多特蒙德市的Uturm场馆前，举行“Interactive Urban Dance Projection”互动体验项目展示

艺术专业与业余的价值观被抛弃。公共空间新媒体艺术中肢体语言表达要求创作者个人或群体的集合性参与，或是和大众欣赏者的群体性合作，运用肢体动作、肢体形态、肢体的表达语言重构新鲜而实在的美术视觉文化和美学观念，应用新手段和新技术条件创造符合社会精神的美术环境，促使对固有审美观的解体，并带来一定的商业机遇。公共空间新媒体艺术的文化艺术与商业价值经营手法不分彼此，艺术也不再肯定个人创作积累和个人英雄成功伟业。公共空间新媒体艺术文化为人们展现的是一个崭新的艺术民主和大众文化时代到来，需要对“多元化”的形式与表现把握更加准确。

2010 年，德国多特蒙德城市曾举办系列的互动体验活动，包括具有实验特色的互动投影，音乐，舞蹈和视觉艺术等。该系列活动隶属于欧洲 Uturm 项目。项目组织众多国际学生和经验丰富的专业互动设计开发团队，所设计的互动作品用于欧洲的文化首都的 E－文化博览会闭幕式。2010 年 8 月 25 日，在德国多特蒙德市的 Uturm 场馆前，举行了“Interactive Urban Dance Projection”互动体验项目展示。该项目将音乐、舞蹈、影像、观众与互动结合，把现场观众的舞蹈动作实时投影至 Uturm 的建筑立面上，并不断增加，最终形成了一面布满抽象、活动的舞蹈影像的立面。

三、结语

现代社会中，由于计算机技术的飞速发展和对人类生活影响程度的日益扩大，新媒体、数字化已渗透到人类生活的各个方面。公共空间与新媒体艺术同时具有“互动”“交流”的社会属性，也有“艺术与文化服务于社会”的精神取向；新媒体艺术的进入，或动态或静态、或是物理存在或是虚拟、再现和表现的艺术形态，形成公共空间新媒体艺术的互动接口，以达到艺术表现、沟通和互动的目的。空间可以是新媒体艺术的媒介，这与新媒体艺术调动一切感官、张扬个性与本能意识的潜在特征相匹配。

公共空间新媒体艺术的叙事表现成为一大趋势并已呈现，公共空间已不再是原有的地域空间，而更发展成以互联网为基础的多维空间，在人类的政治、经济、文化等综合性视角，以多维形态呈现，其作为人类社会生活与接触的重要内容，表现出更强的虚拟性、公共性、时效性及交互性，使公共空间与新媒体艺术的互动在设计和实践中形成交会。

参考文献：

[1]蒋孔阳，德国古典美学[M].北京：商务印书馆，2014：209-211.

（责任编辑：梁 田）

doi: 10.14132/j.cnki.1673-5439.2015.01.003

基于BJND和JNDD的立体视频深度感知增强技术综述

刘 峰[1,2],施 阳[1],干宗良[1],秦 雷[1,2],陈昌红[1]

(1.南京邮电大学 江苏省图像处理与图像通信重点实验室,江苏 南京 210003
2.台湾大学 电机工程学系,台湾 台北 10617)

摘要: 人类视觉系统是一个具有多种视觉特性的复杂处理系统,视觉感知是人眼多种视觉特性共同作用的结果。立体视频图像因视差和深度的存在有着与平面图像不同的视觉感受,且深度感与视差有着紧密的关系。恰可察觉差别(Just Noticeable Difference,JND)或者称为恰可察觉失真(Just Noticeable Distortion,JND)是一种基于视觉心理学和生理学的视觉特性,双目恰可察觉差别(Binocular JND,BJND)和恰可察觉深度差别(Just Noticeable Depth Difference,JNDD)影响人们对立体视频的用户体验。在阐述JND基础上,综述人类视觉感知立体视频的BJND、JNDD模型及国内外研究现状,分析视差图质量和深度图信息对立体视频感知影响,提出了基于JND的深度图处理方法,寻求人眼对立体视频深度感知增强技术,建立基于BJND和JNDD的立体视频图像质量客观评价方法,为立体视频图像处理及应用提供指导。

关键词: 立体视频; BJND; JNDD; 深度感知; 3D图像质量评价
中图分类号: TN919.8 **文献标志码:** A **文章编号:** 1673-5439(2015)01-0026-07

Stereoscopic video depth sensation enhancement technology based on BJND and JNDD

LIU Feng[1,2], SHI Yang[1], GAN Zongliang[1], QIN Lei[1,2], CHEN Changhong[1]

(1. Jiangsu Province Key Lab on Image Processing & Image Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Department of Electrical Engineering, Taiwan University, Taipei 10617, China)

Abstract: Human vision system is a complex processing system with a variety of visual characteristics, the visual perception result from several visual characteristic effects. Stereoscopic video image has distinct in visual perception from 2D image because of the disparity and depth information, and there is a close relationship between the depth sensation and the disparity. Just noticeable difference (JND) is a visual characteristic based on psychology and physiology, the binocular JND (BJND) and just noticeable depth difference (JNDD) can affect users' experiences of the stereoscopic video. Based on introducing JND, this paper summerizes BJND and JNDD models of HVS as well as current research in the international field. By discussing the influences of disparity map quality and depth map information on stereoscopic video perception to improve the stereoscopic video sensation enhancement, a new method for the depth map processing is proposed. The stereoscopic video objective quality assessment model is established based on BJND and JNDD, thus providing the theoretical guidance to the processing and application of the stereoscopic video.

Key words: stereoscopic video; BJND; JNDD; depth sensation; 3D image quality assessment

收稿日期: 2014-12-01 **本刊网址:** http://nyzr.njupt.edu.cn
基金项目: 国家自然科学基金(61471201,61172118)、江苏省高校自然科学研究重大项目(13KJA510004)、江苏省优秀青年教师和校长赴境外研修计划和江苏省"六大人才高峰"高层次人才资助项目
通讯作者: 刘 峰 **电话:** 025-85866737 E-mail: liuf@njupt.edu.cn

随着立体视频技术的快速发展,3D 电影及 3DTV 受到人们越来越多的关注和喜爱,但人们对立体视频的深度感知或体验仍然有待于进一步提高,因为人们观看时需要佩戴立体眼镜,长时间观看会出现视觉疲劳现象等。因此,如何增强人们对立体视频的欣赏舒适度,获得良好的 3D 体验是一个理论界和工业界关注的课题。

立体视频的视觉感知质量增强和用户体验改善可以通过改进多个技术环节来实现。一方面可以加强采集、显示技术等与硬件器件密切相关的物理层面,从根本上减少左右视图信号的采集差异和显示串扰,从而提高视觉质量; 另一方面也可以通过改进立体视频处理中的多项关键技术,减少由多种图像处理对原始三维信号产生的畸变,从而降低视觉失真。

人眼视觉特性是人们在观看视频图像时所反应出来的生理和心理特性,由此建立的恰可察觉差别(Just Noticeable Difference,JND) 模型在平面视频图像处理中,起到重要的指导作用。推广到 3D 领域,双目恰可察觉差别(Binocular Just Noticeable Difference,BJND) 和恰可察觉深度差别(Just Noticeable Depth Difference,JNDD) 是人眼感知立体视频的两个重要概念,在人们增强立体视频真实感体验、提升视频感知质量以及结合人眼视觉特性评价立体视频质量等方面都有着重要的作用,是立体视频图像领域当前研究的热点之一。

本文首先从人类视觉系统(Human Vision System,HVS) 出发分析 JND 模型。其次,结合立体视频分析了 BJND 和 JNDD 模型,综述了当前国际上针对 BJND 和 JNDD 的研究现状和进展,以及在立体视频处理方面的应用,提出了基于 JNDD 的深度图处理方法,以增强人眼对立体视频的深度感知。第三,建立了一种基于 BJND 和 JNDD 的立体视频图像质量客观评价方法。最后,展望了立体视觉特性的未来研究方向。

1　人类视觉特性及 JND

1.1　人类视觉特性

人眼是视频图像的信宿,是视频图像的最终接收对象,人眼作为 HVS 中复杂的器官组织,其观看视频图像的感知或体验受视觉生理和视觉心理的影响。视频图像通过人眼反应到大脑是一个视觉信息处理的过程。目前对人类视觉感知系统包括视觉生理学和视觉心理学的研究,已经取得许多新的进展和突破,如人眼具有的亮度适应能力、对比灵敏度特性、掩蔽效应、深度感知、运动感知等视觉特性。其中掩蔽效应是 HVS 的重要视觉特性之一,是恰可察觉差别的理论基础。该效应主要模拟不同的视觉信号交叉出现在同一个空间区域具有相互影响的视觉现象,如当一个掩蔽信号出现在一个测试信号的背景中时,将会降低(或增强) 图像信号的可视度。

1.2　JND

为了将 HVS 的特性应用到视频图像处理中,需要度量 HVS 对失真的敏感程度,于是提出了 JND 概念。JND 被定义为可见失真的最小值。如果某个信号的 JND 越大,说明人眼对此信号的失真容忍程度也越大,反之,人眼所能允许此信号的失真就越小。

常用的 JND 模型有 3 种分类方法。一是按所依据的 HVS 特性的不同,JND 模型可分为: 背景亮度模型、边缘掩膜模型、对比敏感度函数、运动矢量模型等。二是根据对视频图像数据处理方式的不同可分为像素域 JND 模型[1-6] 和变换域 JND 模型[7-13]。三是有部分学者根据推导模型时考虑的是空间还是时间特性将其分为空域 JND 模型、时域 JND 模型和彩色 JND 模型。其中第二种方式最为常见,像素域是指由像素点组成的二维平面排列的像素采样数据的集合; 变换域是指上述像素域数据以一定的变换规则(如 FFT、DCT、小波变换等) 进行变换得到的结果。

2　立体视频及视觉感知

2.1　立体视频

立体视频通常反映立体场景的信号进入人类双眼,在人类大脑中直接或间接形成具有立体深度的图像信息,使人们欣赏到有身临其境的深度感、真实感和逼真的视觉效果。立体视频系统通常包含采集、处理和显示环节,在实际的立体视频系统中,由于各个处理环节的不当,可能使得最终呈现给用户的双目视频信号有异于日常生活中双眼观看到的自然视觉信号,降低立体效果的体验度。对于这些不自然的双目视频信号,视觉系统需要花费更大的“运算量”进行处理,且有可能发生“处理异常”,因而产生视觉疲劳或不适。

2.2　双目视差及深度图

立体视频的深度感知和体验按其显示的原理不同可分为全息显示、光栅式自由立体显示、基于双目视差的立体显示等种类,其中基于双目视差的双目立体视频显示技术其应用最为广泛。所谓双目视差

是指人双眼有一定瞳距,在观看物体时左眼和右眼所接收到的视觉图像略有差异。基于双目视差的立体显示为观看者的左右眼提供同一时刻同一场景的立体图像对,采用技术手段让观看者的左右眼分别只看到对应的左右眼图像,这样便使观看者感知到立体图像。

深度图表示场景中各点相对于观察者的距离,即深度图中的像素值表示场景中物点与观察者之间的距离。深度图是影响立体视频感知质量重要的一个中间环节,是区别于感知平面图像质量的重要方面。深度图获取最常用的方法是由双目视差图通过空间坐标变换得到深度图像,双目视差为人眼提供具有真实感和深度感的场景信息,但视差过大很容易引起疲劳,视差过小逼真感会减少,从而降低用户的立体感体验。由左右眼视图估计视差从而获取深度图,主要通过视差估计和深度估计得到,由于目前有关视差估计和深度的估计算法仍有待提高,其估计准确性不够高,导致深度图出现失真,如局部的斑点、噪声、空洞及块分割造成的块赝像等,从而直接影响立体视频的感知质量。

2.3 BJND 模型及研究现状

BJND 反映了人类双眼在观看立体视频图像对时刚好可察觉到两者差别的最小失真。或者说,在背景信息和一个视点相应区域的失真给定的情况下,另一个视点能够引起立体图像感知差异的最小失真。BJND 可用来确定立体图像对的失真是否在人眼可察觉的范围之内。在立体图像对中如果其中一个图像的失真小于 BJND,由于双目融合与抑制效应的作用,则人眼无法察觉该立体图像对的失真,反之,人眼非常敏感这些图像点或区域。文献[14]提出基于心理的非对称失真立体图像的 BJND 模型,文中给出了两个实验,实验 1 根据亮度掩盖和双目噪声组合设定了联合阈值;实验 2 对于立体图像,测试了由于对比度掩盖效应而引起的在双目视觉中可视敏感度的降低情况,证明了 BJND 与 HVS 的亮度适应性和对比度掩蔽特点有关,其实验系统如图 1 所示。

在左右视点图像给定的条件下,给出右视点的 BJND 为:

$$BJND_R(bg(i+d),eh(i+d),A_l(i+d)) = A_{C,\text{limit}}(bg(i+d),eh(i+d)) \times \left(1-\left(\frac{A_l(i+d)}{A_{C,\text{limit}}(bg(i+d),eh(i+d))}\right)^{\lambda}\right)^{1/\lambda}$$

其中,$BJND_R$ 表示右视点图像的 BJND,d 表示右视点相对于左视点的视差,$A_{C,\text{limit}}$是考虑对比度掩蔽效应时,左视点随机噪声幅度为零,右视点加入了能够引起双目感知失真的随机噪声幅度上限,$bg(i)$ 是区域 i(例如一个宏块)像素的亮度平均值,参数 λ 控制左视点的噪声影响,其范围为 1.0 ~ 1.5,$eh(i)$ 是区域 i 利用 5 × 5 Sobel 算子得到的边缘梯度,$A_l(i+d)$ 表示左视点相应区域 i 所加的最大可容忍的随机噪声幅度。

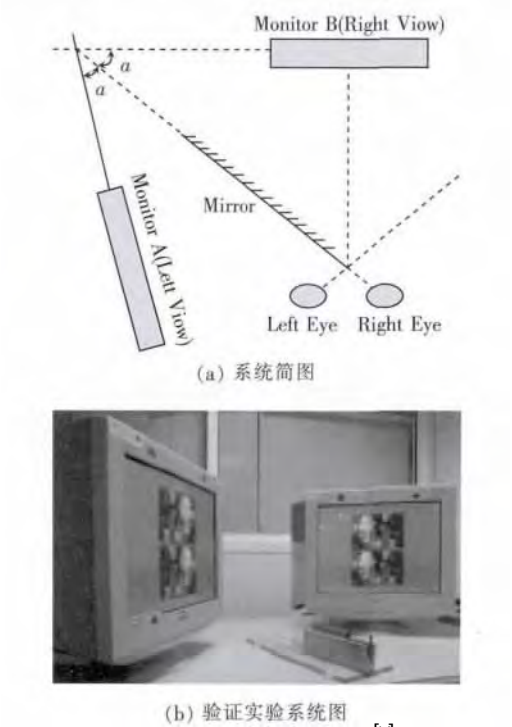


图 1 镜像立体显示系统[14]

BJND 模型在立体视频图像增强和编码、立体视频图像质量评价等方面有着广泛的应用。文献[15]提出了基于自由视点显示器的立体融合系数和聚焦加权模型,并且将其应用到立体 JND 模型中。文献[16]中 Balanna 等人提出了一种基于 BJND 的图像清晰度增强算法,实验结果表明,该算法在一定程度上增强图像的清晰度而不产生超过 BJND 阈值的失真。文献[17]中 Fezza 基于 BJND 和深度图提出了一种能动态决定不对称性最佳边界的数学模型,该方法非均匀地降低了同一场景内立体图像对的分辨率,减少了带宽需求,在非对称编码中有广泛应用。

国内在这方面也有大量研究,如宁波大学郁梅等人研究 BJND 并提出基于 BJND 的快速多参考帧选择算法[18]并应用到多视点立体视频压缩编码中,通过分析 JMVC 多视点视频编码中的多参考帧、双向搜索与 BJND 的统计关系,确定当编码宏块的

BJND 大于阈值时可对多参考帧选择过程进行优化,从而降低编码的复杂度。G. F. Zhu 等人提出了一种基于 BJND 的宏块级码率控制方法[19],这种方法从视点级、图像组级、帧级和码字级分别进行视差匹配,实验证明该方法能更精确的控制码率,得到更好的立体视频主观质量。

2.4　JNDD 模型及研究现状

JNDD 表示人类双眼在一定视距下观察 3D 场景中物体远近变化时,引起人眼深度感知变化的最小阈值,反映了人眼对物体深度感知的能力。通过建立 JNDD 的阈值模型,可以在不降低用户深度感知的前提下对立体视频、立体电影、多视点视频的数据进行有效处理。JNDD 模型与人眼对深度的敏感度特性、双目视差、视网膜模糊和相对尺寸值有关。

文献 [20] 采用自动立体显示实验方法得出观察 3D 立体视频时对应的最小深度容差,反复实验统计拟合出 JNDD 模型,并用此模型对深度图像处理,如图 2 所示。该模型是对人眼所可以感知的最小深度差范围的统计。文中同时指出,人眼对于物体在某一深度向前移动比向后移动视觉感受更加敏感,人眼的 JNDD 阈值与深度有关。

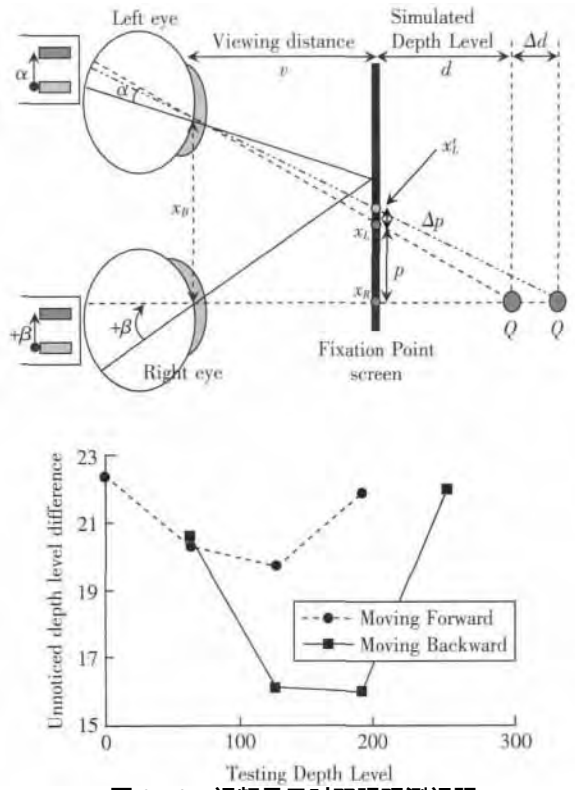


图 2　3D 视频显示时双眼观测视距及 JNDD 与深度变化的关系[20]

文中推导出 JNDD 模型为:

$$D_{JND}(i,j)=\begin{cases}21, & 0\leqslant I(i,j)<64\\ 19, & 64\leqslant I(i,j)<128\\ 18, & 128\leqslant I(i,j)<192\\ 20, & 192\leqslant I(i,j)<255\end{cases}$$

其中,D_{JND} 表示基于 JND 统计的人眼可察觉的最小的深度级容差; $I(i,j)$ 表示深度图中像素点的灰度值。

目前,JNDD 模型的研究主要应用于立体视频图像深度感知增强、视频编码压缩等方面。文献 [21] 中 De Silva 等人建立了一种改进的 JNDD 数学模型,通过这一模型分析了 JNDD 的 3 种深度线索: 双目视差、视网膜模糊和相对尺寸,并经主观质量评价验证,不足之处是该模型实验过程中没有考虑到感兴趣区域影响、模糊图像放置在人眼视觉注意区之外的影响等。文献 [22] 中 Jung 等人对人眼 JNDD 与物体移动深度值关系进行了研究,提出改善人眼观测物体的深度感知方法,根据人眼所能察觉的深度差阈值人为地修改深度图中相邻对象的深度值。主观质量评价说明,该算法能有效改善人的深度感。但不足之处是该方法处理的效果与原始深度图质量以及深度图中的对象分割依赖性太强。

此外,文献 [23] 中 Jung 等人采用修改深度图中邻近物体深度值的差值使其在 JNDD 范围之内,修改方法是由深度数据保留项、深度顺序保留项、深度差分项组成的能量表达式最小,通过立体视频图像的主观评价,深度感知得到有效地增强。不足之处是其性能依赖于原始深度图像及深度图分割效果。

文献 [24] 提出了一种改良的 JNDD 测量方法,这种方法通过调整对象的物理尺寸来维持对象的感知尺寸。在此基础上提出了一种深度自适应裂痕修补技术,用来对裂痕区域进行高精度的补偿。实验结果证明了该方法的有效性。从目前检索的资料来看,De Silva [25-26] 及 Jung 是当前在该领域研究成果最为突出的两位学者。

国内也有不少研究者开始关注 JNDD 对立体视频图像深度变化感知的影响,文献 [27] 通过主观实验得出了 JNDD 模型,并提出了一种基于 JNDD 模型的双边滤波深度图像处理优化方法,可较明显地提高深度图的质量,并在减少深度图编码码率的情况下,提高合成虚拟视点的质量。文献 [28] 提出了一种基于 JND 深度阈值的立体图像处理新方法,首先将立体图像分割成遮蔽区域和非重叠区域,然后对不同的区域用相应的深度阈值处理。考虑到深度阈值的作用,文献 [29] 在研究一种新的 MJND(多视点中的 JND) 模

型时,也包含了类似于 JNDD 的 DJND(深度 JND)部分,有效提高了立体视频的压缩效果。

2.5 深度感知

深度感知是指人眼辨别明显发生前后位移的物体相对距离的能力。实际上,人眼在观看具有深度变化物体或立体显示时,只能感知大于 JNDD 值的深度变化,所有小于 JNDD 的深度变化都被人眼所忽略。由于获取深度图过程中,所得深度图的局部斑点、噪声与深度图中实际的边缘信息是有差异的,于是在某些情况下,存在着与实际边缘信息的像素差值小于 JNDD 的许多像素点,可以通过应用 JNDD 模型将这种噪声删除掉。

因此,利用人眼视觉系统的 BJND 和 JNDD 模型,结合立体视频图像对的视差图、深度图等与深度感知密切相关的估计和处理,提出了立体视频感知质量增强的深度图像处理方法,流程如图 3 所示。针对传统方法在立体匹配过程中匹配算法存在目标边缘及精细纹理区产生匹配误差、匹配不准确问题,引入视觉阈值 JNDD 模型,根据人眼视觉对立体图像的深度感知,对深度图的像素点灰度级进行调整,并对深度图像坏点进行剔除、对空洞进行填补,得到基于 HVS 特性的深度增强图像。

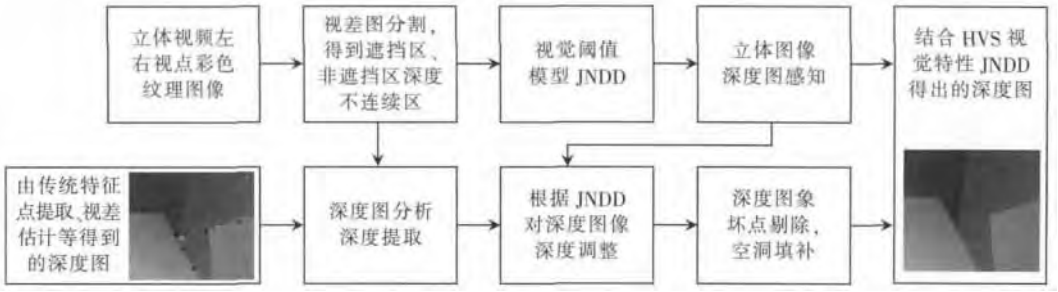


图 3 立体视频深度图像处理

3 立体视频质量客观评价

3.1 立体视频质量评价概述

立体视频质量评价是立体视频系统的非常内容,视频处理系统的整体性能最终反映到观看者对立体视频的切身体验上。立体视频质量评价分为主观质量评价(SQA)和客观质量评价(OQA)[30]。主观质量评价是让大量测试者按照事先规定的评价规则对测试图像按视觉效果的优劣进行评分,对所有测试者评分得到的原始数据进行处理得到实验的有效数据。客观质量评价方法是通过数学建模等手段设计出一套计算模型,自动地对测试图像进行评分,然后通过与主观质量评价之间的相关性来验证该模型的正确性和可靠性。对于立体视频图像质量客观评价由于受视差及深度感知的因素影响,将传统的平面图像质量客观评价方法用于立体视频并不合适[31-34]。因此,需要研究结合人眼视觉特性的立体视频质量的客观评价方法。

3.2 基于 BJND 和 JNDD 的立体视频质量客观评价方法

目前关于立体视频质量的评价研究如火如荼,如文献[35]中 Hachicha 等人在研究立体图像质量评价时引入了一种 BJND 模型,一方面计算了参考立体图像左右视点间 JND,同时也模仿了双目抑制理论。实验结果表明,该模型在不考虑非对称失真的情况下是有效的。针对人眼双目视觉系统具有亮度对比度敏感性、时空掩蔽、双目掩蔽等特性。文献[36]中 Qi 等人参照这些特性提出了一种类似 BJND 的模型,并基于此双目感知模型改良获得了一种全参考立体视频质量评价方法,实验证明这种方法与主观感知具有较好一致性,但该模型没有考虑立体视频的某些统计特性。上海大学张艳等人[37]研究了双目立体视频最小可察觉差别模型,并将其应用在图像质量评价。然而,此方法主要建立在特定的实验条件基础上,对于实际应用有一定的局限性。邵枫等人[38]针对于立体视频不同于单路视频的特性建立了考虑多种视觉特性因素的 BJND 模型,并将该模型应用于立体视频质量的客观评价中。

本文在综述并分析上述基于人眼视觉特性的立体视频质量评价基础上,建立了一种结合 BJND 和 JNDD 模型的立体视频图像的客观评价方法,使得立体图像质量的客观评价尽可能与人眼主观视觉评价一致,其评价模型如图 4 所示。关键是在原始左右视点图像、失真左右视点图像、失真前后左视点图像、失真前后右视点图像分析过程中,充分利用 HVS 特性如图像结构相似度、掩蔽现象、BJND、

JNDD 等现象,以求更加准确的反映人眼视觉特征。

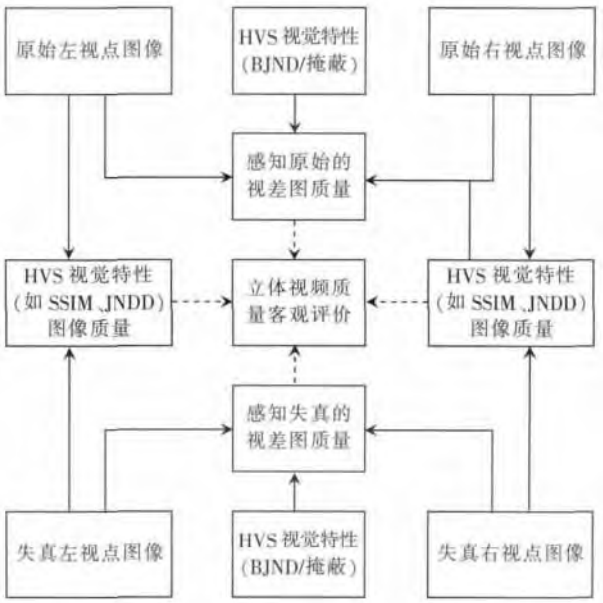


图 4　基于视觉感知模型的立体视频图像质量客观评价方法示意图

4　结束语

相比 JND 技术,BJND 和 JNDD 的研究还相对较少,但随着立体视频的不断发展和需求的不断增长,BJND 和 JNDD 将发挥越来越重要的作用,尤其是在以下方面仍然需要深入研究: (1) 无论是 JND 还是 BJND、JNDD,都与人眼视觉特性密切相关,而人眼在观看立体视频所表现的视觉现象如立体感体验、视觉疲劳、显著性注意区域等仍有许多值得进一步探索的课题。因此,随着人眼立体视觉特性研究的不断深入,必将可以得到更加符合人眼特性的 BJND 和 JNDD 数学模型。(2) 人眼视觉系统的各个部分相互影响且处理机制非常复杂,类似于人眼存在着时间频率响应、空间频率响应以及两者相结合的时空频率响应一样,在研究人眼的立体视觉特性时,结合多种视觉特性模型如 BJND 和 JNDD 两者结合的联合模型是后续需要研究的重要内容。(3) 目前,基于 HVS 的视觉特性不断地引入到 2D 图像质量的客观评价中,如 SSIM (Structural SIMilarity)、VQM (Video Quality Metric) 等,但由于立体视频的双目特性和深度特性,使其无法复制原有的评价方法,结合立体视觉特性的立体视频质量客观评价,有待进一步测试和验证。

参考文献:

[1] YANG X K, LING W S, LU Z K, et al. Just noticeable distortion model and its applications in video coding [J]. Signal Processing: Image Communication, 2005, 20(7): 662 – 680.

[2] HUANG T H, LIANG C K, YEH S L. JND-Based enhancement of perceptibility for dim image [C] // IEEE Conference of International Image Communication and Image Processing. 2008: 1752 – 1755.

[3] HUANG T H, KAO C T, CHEN I C, et al. A visibility model for quality assessment of dimmed images [C] // Fourth International Workshop Quality Multimedia Experience. Yarra Valley, Australia, 2012.

[4] HUANG T H, KAO C T, CHEN H H. Quality assessment of images illuminated by dim LCD backlight [C] // Proc Human Vision Electronic Imaging XVII. 2012.

[5] YANG X K, LIN W S, LU Z. Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images [C] // IEEE International Conference on Acoustics, Speech, and Signal Processing. Hongkong: IEEE Press, 2003: 609 – 612.

[6] LIU Anmin, LIN Weisi, ZHANG Fan, et al. Enhanced Just Noticeable Difference (JND) estimation with image decomposition [C] // 17th IEEE Intemational Conference on Image Proeessing. Hongkong: IEEE Press, 2010: 317 – 320.

[7] JIA Y T, LIN W, KASSIM A A. Estimating just-noticeable distortion for video [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(7): 820 – 829.

[8] ZHANG X H, LIN W S, XUE P. Just-noticeable difference estimation with pixels in images [J]. Journal of Visual Communication and Image Representation, 2008, 19(1): 30 – 41.

[9] CHEN Z Z, GUILLEMOT C. Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(6): 806 – 819.

[10] CHEN Z Z, GUILLEMOT C. Perception-oriented video coding based on foveated JND model [C] // Picture Coding Symposium. Chicago, USA, 2009: 1 – 4.

[11] ZHANG X, LIN W S, XUE P. Improved estimation for just-noticeable visual distortion [J]. Signal Processing, 2005, 84(4): 795 – 808.

[12] WEI Z Y, NGAN K N. Spatial just noticeable distortion profile for image in DCT domain [C] // IEEE International Conference on Multimedia and Expo. Hannover: IEEE Press, 2008: 925 – 928.

[13] ZHANG X, LIN W S, XUE P. Just-Noticeable difference estimation with pixels in images [J]. Journal of Visual Communication and Image Representation, 2008(19): 30 – 41.

[14] ZHAO Y, CHEN Z Z, ZHU C, et al. Binocular just-noticeable-difference model for stereoscopic images [J]. IEEE Signal Process Letters, 2011, 18(1): 19 – 22.

[15] ZHANG L, PENG Q, WANG Q H, et al. Stereoscopic perceptual video coding based on just-noticeable-distortion profile [J]. IEEE Transactions on Broadcasting,

2011,57(2): 572 - 581.

[16] BALANNA P,SUVARNA K,SUDHAKAR K. Improved depth conception with sharpness augmentation for stereo video [C] //IJCER. 2013.

[17] FEZZA S A,LARABI M C,FARAOUN K M. Asymmetric coding using Binocular Just Noticeable Difference and depth information for stereoscopic 3D [C] //IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP). 2014: 800 - 884.

[18] ZHOU Junming,JIANG Gangyi,YU Mei. Subjective quality analyses of stereoscopic images in 3DTV system [C] // Visual Communications and Image Processing (VCIP). 2011: 1 - 4.

[19] ZHU G F,YU M,JIANG G Y. A novel macroblock level rate control method for stereo video coding [J/OL]. [2014-11-03]. http://dx.doi.org/10.1155/2014/136854.

[20] DE SILVA D, EKMEKCIOGLU E, FERNANDO W, et al. Display dependant of depth maps based on preprocessing just noticeable depth difference modeling [J]. IEEE Transactions on Signal Processing: Selected Topics in Signal Processing, 2011,5(2): 335 - 351.

[21] DE SILVA D, FERNANDO A, WORRALL S, et al. Sensitivity analysis of the human visual system for depth cues in stereoscopic 3-D displays [J]. IEEE Transactions on Multimedia, 2011,13(3): 498 - 506.

[22] JUNG S W, KO S J. Depth sensation enhancement using the Just Noticeable Depth Difference [J]. IEEE Transactions on Image Processing, 2012, 21(8): 1191 - 1199.

[23] JUNG S W,KO S J. Depth enhancement considering just noticeable difference in depth [J]. IEICE Transactions on Fundam Electron,2012,95(3): 673 - 675.

[24] JUNG S W. A modified model of the just noticeable depth difference and its application to depth sensation enhancement [J]. IEEE Transactions on Image Processing, 2013, 22(10): 3892 - 3903.

[25] DE SILVA D V S X, FERNANDO W A C, NUR G, et al. 3D video assessment with just noticeable difference in depth evaluation [C] //7th IEEE International Conference on Image Processing. 2010: 4013 - 4016.

[26] DE SILVA D V S X,FERNANDO W A C,WORRALL S T, et al. Just noticeable difference in depth model for stereoscopic 3D displays [C] //IEEE International Conference on Multimedia and Expo. 2010: 1219 - 1224.

[27] LIU Xingyu,YU Mei,TIAN Tao. New just-noticeable-distortion model based on the depth information and its application in multiview video coding [C] // Applied Mechanics and Materials. 2013: 2512 - 2517.

[28] LI Xiaoming,WANG Yue,ZHAO Debin. Joint just noticeable difference model based on depth perception for stereoscopic images [C] // Visual Communications and Image Processing (VCIP). 2011: 1 - 4.

[29] ZHOU Lili,WU Gang,HE Yan. A new just-noticeable-distortion model combined with the depth information and its application in multi-view video coding [C] //Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). 2012: 246 - 251.

[30] SESHADRINATHAN K,SOUNDARARAJAN R,BOVIK A C, et al. Study of subjective and objective quality assessment of video [J]. IEEE Transactions on Image Processing, 2009,19(6): 1427 - 1441.

[31] HEWAGE C T E R, MARTINI M G. Quality evaluation for real-time 3D video services [C] //IEEE International Conference on Multimedia and Expo (ICME). 2011: 1 - 5,11 - 15.

[32] EL-YAMANY N A, UGUR K, HANNUKSELA M M, et al. Evaluation of depth compression and view synthesis distortions in multiview-video-plus-depth coding system [C] // The True Vision-Capture, Transmission and Display of 3D Video (3DTV'11). Antalya, Turkey,2011: 1 - 4.

[33] AKHTER R, PARVEZ SAZZAD Z M, HORITA Y,et al. No-reference stereoscopic image quality assessment [C] //Proc of SPIE-IS&T Electronic Imaging. 2010.

[34] MAALOUF A, LARABI M C. A Stereo Color Image Quality Assessment Metric [C] //IEEE International Conference on Computer Science and Automation Engineering. 2011.

[35] WALID H,AZEDDINE B,ALAYA C F. Stereo image quality assessment using a binocular just noticeable difference model [C] //20th IEEE International Conference on Image Processing. 2013: 1191 - 1199.

[36] QI F,JIANG T,FAN X. Stereoscopic video quality assessment based on stereo just noticeable difference model [C] //20th IEEE International Conference on Image Processing. 2013: 34 - 38.

[37] 张艳, 安平, 张秋闻, 等. 双目立体视频最小可辨失真模型及其在质量评价中的应用 [J]. 电子与信息学报, 2012,34(3): 693 - 703.

[38] SHAO F,LIN W,GU S. Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics [J]. IEEE Transactions on Image Processing, 22(5): 1940 - 1953.

作者简介:

刘 峰(1964 -),男,浙江文成人。南京邮电大学教育科学与技术学院副院长,江苏省图像处理与图像通信重点实验室教授,博士生导师。研究方向为图像处理与多媒体通信、虚拟现实。

基于 MapReduce 的 SVM 分类算法研究

秦 军[1],戴新华[2],童 毅[2],林巧民[1]
(1. 南京邮电大学 教育科学与技术学院,江苏 南京 210003;
2. 南京邮电大学 计算机学院,江苏 南京 210003)

摘 要: 云计算环境中,传统的基于 MapReduce 的 SVM 分类算法对数据集的训练是将各子节点训练后得到的支持向量进行合并,得到的分类器分类效率和准确率不理想。为此,文中提出了一种改进的训练算法,在各节点上运用遗传算法来寻找子数据集的最优核函数及参数,用得到的参数组合对子数据集进行训练得到支持向量,合并每个节点训练后的支持向量为全局支持向量,然后在各个节点上将子集与全局支持向量合并作为新的训练数据集。重复这四个步骤,直到全局支持向量不再变化时,则收敛到最优分类模型。最后,经开源云计算平台 Hadoop 实验验证,该算法的分类正确率比传统的分类算法有了明显提高。

关键词: MapReduce; SVM 分类算法; 遗传算法; 云计算
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 1673-629X(2015)06-0087-05
doi: 10.3969/j.issn.1673-629X.2015.06.019

Research on SVM Classification Algorithm Based on MapReduce

QIN Jun[1],DAI Xin-hua[2],TONG Yi[2],LIN Qiao-min[1]
(1. College of Education Science & Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. College of Computer,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract: In cloud computing environment,the method adopted by the traditional SVM sorting algorithms based on MapReduce of training data set is too simple and it just merges support vectors after nodes' training,so the efficiency and accuracy of classifier are not very ideal. To solve the problem above,an improved training algorithm is proposed in this paper. Firstly,use the genetic algorithm to get the optimal kernel function and parameters on each node at the same time,then using the combination to train the data set for support vector,and afterwards,combining all support vectors after training as a global support vector,and then merging every data subset with global support vector on each node to get a new training data set. Repeat these four steps until the global support vector no longer changes and that's to say,it converges to the optimal classification model. Finally,the experiment on Hadoop proves that the classification accuracy of new algorithm is improved obviously than traditional classification algorithms.

Key words: MapReduce; SVM sorting algorithm; genetic algorithm; cloud computing

0 引 言

随着网络技术的快速发展,各类信息数据的增长速度也越来越快。如何高效而快速地在海量数据中找到用户需要的信息,是当今信息处理技术的一个难题[1]。

传统的单机平台由于存储及计算能力的约束,对大量数据的分析处理效率低下。为了提高分类质量与效率,首先要实现算法的并行化。

在机器学习领域,常见的分类算法有 K 最近邻(KNN)、贝叶斯(Bayes)、神经网络、决策树和支持向量机(SVM)。其中,支持向量机算法由于其广义属性,可提供非常强大的、精确的分类方法[2]。但算法在计算和存储中数据集的不断增加,支持向量机在内存需求和计算时间上达到了瓶颈状态。基于 MapReduce[3] 的支持向量机算法克服了大规模数据集的制约,扩展了 SVM 算法的功能。但传统算法只是对各子节点的数据子集进行训练后得到的支持向量简单地进行合并得到分类器[4],用数据集测试后发现,各子节

收稿日期: 2014-07-22 **修回日期:** 2014-10-28 **网络出版时间:** 2015-05-06
基金项目: 江苏省自然科学基金项目(BK20130882)
作者简介: 秦 军(1955-),女,教授,研究方向为计算机网络技术、多媒体技术、数据库技术。
网络出版地址: http://www.cnki.net/kcms/detail/61.1450.TP.20150506.1621.004.html

点上的分类准确率较低。由此,文中提出了改进的基于 MapReduce 的 SVM 分类算法。

1 MapReduce 编程模型与应用

MapReduce 技术是由 Google 公司在 2004 年最先提出的,作为面向大数据分析和处理的并行计算模型,MapReduce 由于其良好的可扩展性、可用性、容错性,成为数据库领域的研究热点。为处理海量数据提供了很好的数据存储和数据处理方法[5]。

1.1 MapReduce 编程模型

MapReduce 是 Map 函数和 Reduce 函数相组合的编程模型[6]。MapReduce 将数据处理分成 Map 和 Reduce 阶段。在 Map 阶段,每个 Map 任务从全局文件系统或数据库中读取一个数据片段,使用用户自定义的 Map 函数将输入数据解析成中间 key/value 对,进行排序和分区后存储在本地;在 Reduce 阶段,Reduce 任务从每个 Map 任务节点读取相应分区数据,使用用户定义的 Reduce 函数对中间 key/value 对进行处理,将结果存储在全局文件系统或数据库中[7]。

MapReduce 操作的相关类型如下表示:

$$Map(key_{in}, value_{in}) \rightarrow list(key_{out}, value_{intermediate})$$

$$Reduce(key_{out}, list(value_{intermediate})) \rightarrow list(value_{out})$$

1.2 传统的基于 MapReduce 的 SVM 算法的分类模型训练

传统的分类算法利用 Map 和 Reduce 函数在多个子节点上进行并行数据处理[8]。首先在 Map 函数上找出所有子训练数据集里面的子支持向量(Support Vectors,SVs),然后利用 Reduce 函数将各个子节点上的子支持向量进行合并,汇总成所有数据集的完整的支持向量 AllSVs。AllSVs 将确定数据集的分类超平面,即得到最终的分类器。

1.3 基于遗传函数对 SVM 核函数及参数的选取

SVM 算法中最大的特点是使用核函数将训练样本映射到特征空间进行分类,训练的效果取决于两个参数:一个是核函数 K 及其参数(RBF 参数为 gama),一个是惩罚因子 C。这两个参数作用于训练过程中的分类函数,对支持向量机的训练结果将产生直接的影响。如果选择不当,会直接导致分类器的效率降低。因此选取好的核函数与参数将提升分类器的性能[9]。

遗传算法是计算机科学人工智能领域中用于解决最优化的一种搜索启发式算法,是进化算法的一种。这种启发式通常用来生成有用的解决方案来优化和搜索问题。遗传算法已被广泛地应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域[10]。利用遗传算法对 SVM 模型选出最优核函数及其参数与惩罚因子。

遗传算法随机生成多个种群个体,将种群中各个体的基因串解码为相应核函数编号、核函数参数和错误惩罚因子,将这三个参数代入 SVM,以训练数据和测试数据对其进行训练和测试,然后计算每个个体的适应度,如果满足条件,则停止,否则重新循环[11]。具体流程如图 1 所示。

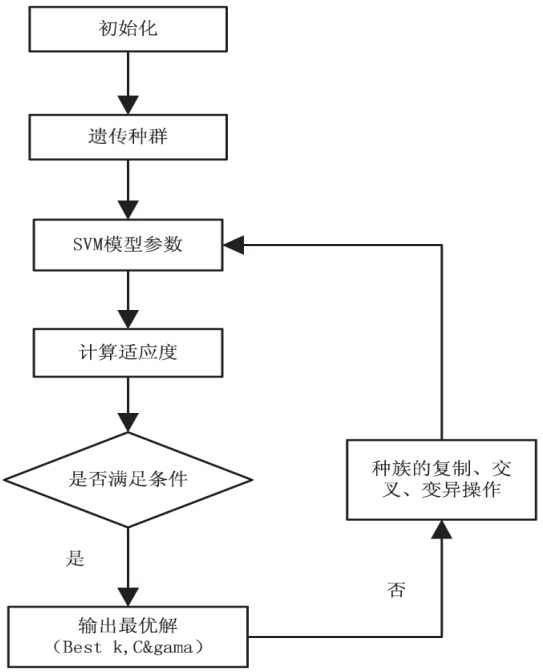


图 1 遗传算法对 SVM 参数的优化流程

1.4 新型的基于 MapReduce 的 SVM 算法的分类模型训练

传统的算法对各子节点的数据子集进行训练后得到的支持向量只是简单地进行合并,得到分类器,用数据测试后发现,各子节点上的准确率都不算高。由此,提出了改进算法:基于 MapReduce 的并行迭代 SVM 算法(Parallel Iterative SVM Algorithm based on MapReduce,PISVMAM)。该算法首先将合并后的支持向量再输入到子节点进行合并[12],运用遗传算法得到最优的核函数及其参数与惩罚因子,然后运用此参数组合对数据集进行训练得到新的子支持向量,再合并,当迭代到全局支持向量不再改变时,迭代结束,得到最终的分类超平面。SD 代表训练数据集在各节点的子集,SVG 为全局支持变量。具体流程如图 2 所示。

2 基于 MapReduce 的并行迭代 SVM 算法

PISVMAM 运行在集群上,将训练集分成各个子集,并分布到各个节点上进行训练,以获得各自的子支持向量。在 MapReduce 作业的 Map 阶段,各节点的子训练集与全局支持向量集合并,然后用遗传算法对合并后的数据集进行训练和测试以选取最优的核函数及

其参数与惩罚因子; 在 Reduce 阶段,运用得到的参数组合对数据集进行训练得到子支持向量。之后各个节点的新的支持向量再与全局支持向量合并成为新的全局支持向量。当迭代多次之后,全局支持向量不再改变时,则算法收敛,得到最优的分类超平面,即最优分类器。

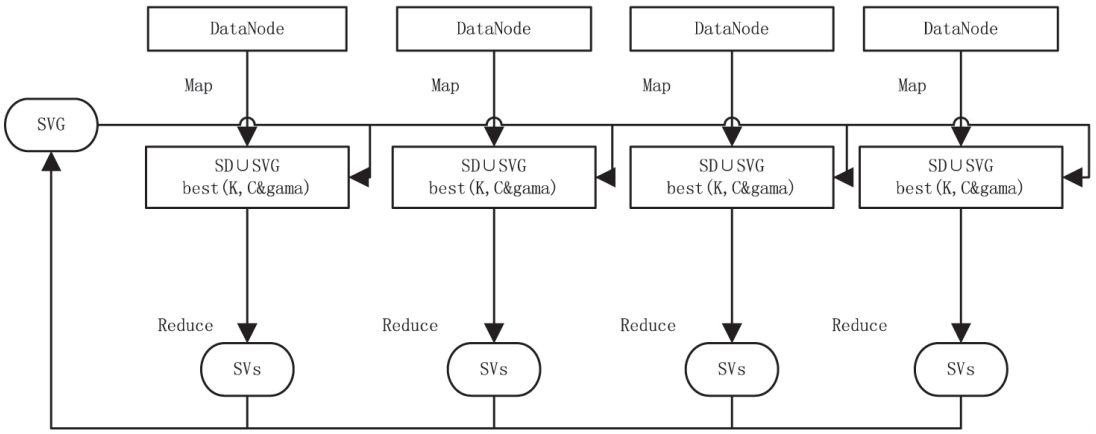


图 2　改进后的算法训练流程

2.1　算法收敛分析

支持向量机的监督学习实际上就是一个经验风险或者结构风险函数的最优化问题。风险函数度量平均意义下模型预测的好坏,模型每一次预测的好坏用损失函数来度量。它从假设空间 F 中选择模型 f 作为决策函数,对于给定的输入 X,由 $f(X)$ 给出相应的输出 Y,这个输出的预测值 $f(X)$ 与真实值 Y 可能一致也可能不一致,用一个损失函数来度量预测错误的程度。损失函数记为 $L(Y,f(X))$。

选取链式损失函数来检验被算法训练的模型。因为链式损失函数使得越偏离边界,受到的惩罚越高。链式损失函数的表达式为:

$$L(f(x),y) = \max\{0,1 - y.f(x)\}$$

模型 $f(X)$ 关于训练数据集的平均损失成为经验风险,表示为:

$$R_{\text{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i))$$

经验风险最小化策略认为,经验风险最小的模型就是最优模型,按照经验风险最小化求最优模型就是求解如下最优化问题:

$$\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i))$$

让 SVG^s 表示经过 s 次迭代后的全局支持向量,节点 l 经过 s 次迭代之后的模型为 f_l^s。

由 SMO 算法[13]:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j < x_i,x_j >$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,n, \sum_{i=1}^{n}\alpha_i y_i = 0$$

当模型由低维映射到高维时,式中 x_i 转换为 $\varphi(x_i)$,引入核函数,则 $< \varphi(x_i),\varphi(x_j) >$ 转化为 $k(x_i,x_j)$,求拉格朗日参数 $\boldsymbol{\alpha}$ 即为求分类模型 f。则

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} Q\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,n, \sum_{i=1}^{n}\alpha_i y_i = 0$$

分解算法把乘子划分成工作集 B 和非工作集 I,由分类算法优化问题变为:

$$\max_{\boldsymbol{\alpha}} [\boldsymbol{e}_B,\boldsymbol{e}_I]\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_I\end{bmatrix} - \frac{1}{2}[\boldsymbol{\alpha}_B,\boldsymbol{\alpha}_I]\begin{bmatrix}\boldsymbol{Q}_{BB} & \boldsymbol{Q}_{BI}\\ \boldsymbol{Q}_{IB} & \boldsymbol{Q}_{II}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_I\end{bmatrix}$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C, i = 1,2,\cdots,n, \sum_{i=1}^{n}\alpha_i y_i = 0$$

由分类算法的收敛性得上式收敛。假设算法在第 s 次迭代之后收敛,则在节点 l 上 α 有最大值,f_l 为最优分类函数。则 $R_{\text{emp}}(f_l^{s+1}) = R_{\text{emp}}(f_l^s)$,$l$ 节点上的支持向量不再变化,$\text{SV}_l^{s+1} = \text{SV}_l^s$,而 $\text{SVG}^{s+1} = \text{SVG}^s \cup \{\text{SV}_l^{s+1} \mid l = 1,2,\cdots,n\}$,则 $\text{SVG}^{s+1} = \text{SVG}^s$,算法收敛,得到最优分类器。

2.2　算法的实现过程

PISVMAM 首先在分布式系统的每台计算机上先读取全局支持向量,然后将本地子训练数据集与全局支持向量合并,对合并后的数据集利用 GA 进行训练与测试选取最优的参数组合,然后再用此参数组合训练数据集,待各个节点训练完成后,再将各个节点的支持向量与全局向量合并。如此循环,当全局支持向量不再变化时,迭代结束。s 为迭代次数,SVG 为全局支持向量,n 为参与训练的子节点数,SVG^s 为迭代 s 次之后的全局支持向量,f^s 为迭代 s 次之后的分类函数。具体实现步骤如下:

步骤 1: 初始化全局支持向量,迭代次数 $s = 0$,全局支持向量 $\text{SVG} = \varnothing$。

步骤 2: $s = s + 1$,在 n 个节点上同时读取全局支持向量集 SVG 与本机器上的训练数据合并得到新的数据集。

步骤 3: 运用遗传算法对新的数据集进行训练与测试,得到最优的核函数 K ,参数 γ ,惩罚因子 C 。

步骤 4: 在所有节点上运用各自的最优参数组合对数据集进行训练,分别得到新的支持向量。

步骤 5: 当每台机器都完成了训练,将所有节点上的新的支持向量与全局支持向量进行合并,得到新的全局支持向量。

步骤 6: 当 $SVG^s = SVG^{s-1}$,则停止,得到最优分类函数 f ,否则重复步骤 2 ~ 6。

3 实 验

LibSVM [14] 是台湾大学林智仁副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包,提供了编译好的可执行文件、源代码,方便改进使用。该软件对 SVM 所涉及的参数调节相对较少,提供了很多的默认参数,利用这些参数可以解决很多问题,并提供了交互检验的功能。文中将使用 libSVM 软件包对实验数据进行训练和测试。

3.1 实验环境及数据集

实验使用 4 台计算机组成集群,4 个节点分别为 1 个 namenode 和 3 个 datanode,每台机器 CPU 为 Intel Pentinum(R) 双核 T4200 2. 0 GHz,,操作系统均为 Ubuntu11. 10 版本,JDK 版本为 JDK_1. 6. 0_45,Hadoop [15] 版本为 1. 0. 1,实验中使用 libSVM 软件包。

实验采用的数据集来自 UCI 数据库,分别为 Spambase 与 eeg eye state 数据集。Spambase 共有 57 个特征,有 4 601 个数据样本,取 3 450 个作为训练数据样本,剩下的 1 151 个作为测试数据样本。同样数据集 eeg eye state 共有 15 个特征,选取 8 000 个数据样本,取 6 000 个作为训练数据样本,剩下的 2 000 个作为测试数据样本。

3.2 实验结果与分析

在对 Spambase 数据集进行训练时,三个节点平均分配 1 150 个作为子训练数据集,对 eeg eye state 数据集进行训练时,每个节点取 2 000 个作为子训练数据集。遗传算法中以分类器的准确率作为适应度函数。在实验过程中两个实验数据集的全局支持向量与所有节点迭代数次之后的收敛图如图 3 ~ 6 所示。

由图 3、图 5 得出 Spambase,egg 数据集分别迭代 3 次和 4 次后就达到了稳定状态,由图 4、图 6 得出分类的准确率也随之不断提高直至达到最大值,再提高迭代次数,全局支持向量与分类准确率无任何变化。通过测试,用传统的基于 MapReduce 的 SVM 方法训练 Spambase 与 egg 数据集,最后得到的平均测试准确率分别为 86. 1% 与 82. 3% 。图 4、图 6 所得 Spambase、egg 数据集的最终分类平均准确率与传统的算法相比有了很大提升。

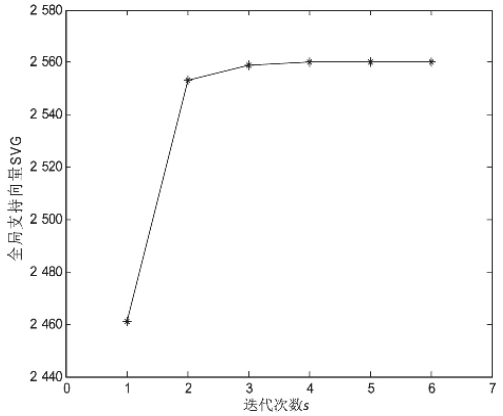


图 3 Spambase 数据集 SVG 收敛图

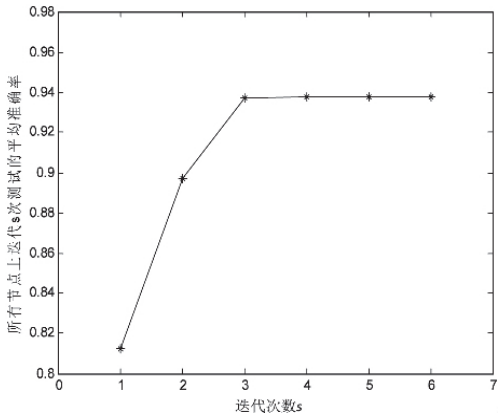


图 4 Spambase 数据集测试准确率收敛图

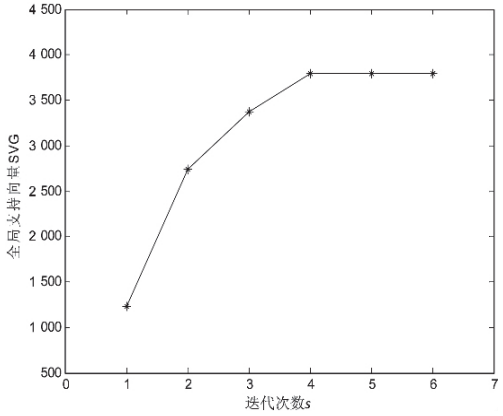


图 5 eeg 数据集 SVG 收敛图

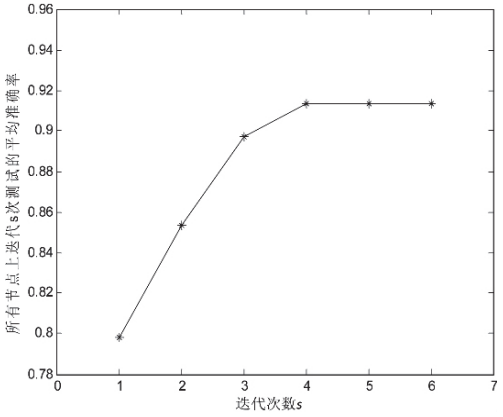


图 6 eeg 数据集测试准确率收敛图

4　结束语

针对传统的基于 MapReduce 的 SVM 分类算法在各子节点上的分类准确率的问题,文中提出了 PISVMAM。该算法引入了迭代机制和遗传算法对 SVM 参数进行优化,将遗传算法运用在每一次迭代中,克服了人为选取 SVM 参数的随意性,使得支持向量机在每一次迭代中都能得到最佳参数,从而在收敛时,得到最优的分类器。实验数据表明,PISVMAM 得到的分类器与传统算法相比,分类准确率有了很大提高。该算法对于 SVM 分类算法来说,能够很好地对海量数据集进行训练和分类。

参考文献:

[1] Chen Kang,Zheng Weimin. Cloud computing system instances and current research [J]. Journal of Software,2009,20(5):1337 – 1348.

[2] Theodoridis S,Koutroumbas K. 模式识别 [M]. 北京: 电子工业出版社,2010.

[3] 黄　山,王波涛,王国仁,等. MapReduce 优化技术综述 [J]. 计算机科学与探索,2013,7(10):865 – 885.

[4] White T. Hadoop 权威指南 [M]. 周敏奇,王晓玲,金澈清,等,译. 北京: 清华大学出版社,2011.

[5] 赵保学,李战怀,陈　群,等. 基于共享的 MapReduce 多查询优化技术 [J]. 计算机应用研究,2013,30(5):1405 – 1409.

[6] 刘　鹏,黄宜华,陈卫卫. 实战 Hadoop [M]. 北京: 电子工业出版社,2011.

[7] 周　锋,李旭伟. 一种改进的 MapReduce 并行编程模型 [J]. 科协论坛: 下半月,2009(2):65 – 66.

[8] 廖周宇,谢晓兰,刘建明. 云计算环境下基于 SVM 的数据分类 [J]. 桂林理工大学学报,2013,33(4):765 – 769.

[9] 奉国和. SVM 分类核函数及参数选择比较 [J]. 计算机工程与应用,2011,47(3):123 – 124.

[10] 马永杰,云文霞. 遗传算法研究进展 [J]. 计算机应用研究,2012,29(4):1201 – 1206.

[11] 刘鳍洁,陈桂明,刘小方,等. 基于遗传算法的 SVM 参数组合优化 [J]. 计算机应用与软件,2012,29(4):94 – 96.

[12] Catak F O,Balaban M E. CloudSVM: training an SVM classifier in cloud computing systems [M]. Berlin: Springer – Verlag,2013.

[13] JerryLead. SMO 算法的数学推导 [EB/OL]. 2011 – 01 – 03. http://www.cnblogs.com/jerrylead/aechive/2011/03/18/1988419.html.

[14] Chang Chih – Chung,Lin Chih – Jen. LIBSVM: a libraray for support vector machine [J]. ACM Transactions on Intelligent Systems and Technology,2011(2):1 – 27.

[15] Apache Hadoop [EB/OL]. 2012 – 04 – 16. http://hadoop.apache.org/.

(上接第 86 页)

电路设计 [J]. 电子器件,2008,31(4):1197 – 1200.

[4] Gu J,Wu J,Gu D,et al. All – digital wide range precharge logic 50% duty cycle corrector [J]. IEEE Trans on Very Large Scale Integration Systems,2012,20(4):760 – 764.

[5] Min Y J,Jeong C H,Kim K Y,et al. A 0.31 – 1 GHz fast – corrected duty – cycle corrector with successive approximation register for DDR DRAM application [J]. IEEE Trans on Very Large Scale Integration Systems,2012,20(8):1524 – 1528.

[6] Chung C C,Sheng D,Shen S E. High – resolution all – digital duty – cycle corrector in 65 – nm CMOS technology [J]. IEEE Trans on Very Large Scale Integration Systems,2014,22(5):1096 – 1105.

[7] Yun W J,Lee H W,Shin D,et al. A 3.57Gb/s/pin loe jitter all – digital DLL with dual DCC circuit for GDDR3 DRAM in 54 – nm CMOS technology [J]. IEEE Trans on Circuits Syst II, Exp Briefs,2011,19(9):1718 – 1722.

[8] 杜振场,殷　勤,吴建辉,等. 一种固定下降沿的高精度时钟占空比调整电路 [J]. 微电子学,2007,37(5):739 – 743.

[9] 张炜华,姚若河,吴桐庆. 一种改进的模拟占空比矫正电路 [J]. 微电子学与计算机,2007,24(3):174 – 177.

[10] 张炜华,姚若河,朱建培. 一种新型的模拟占空比矫正电路 [C]//第十四届全国半导体集成电路、硅材料学术年会. 北京: 出版者不详,2005.

[11] 李　华,钟　正,方　粮,等. 占空比调节器的设计与实现 [C]//第十二届计算机工程与工艺全国学术年会. 呼和浩特: 出版者不详,2008.

[12] Cheng K,Su C,Chang K. A high linearity,fast – locking pulse width control loop with digitally programmable duty cycle correction for wide range operation [J]. IEEE Journal of Solid – state Circuits,2008,43(2):399 – 413.

[13] Han S,Kim J. Hybrid duty – cycle corrector circuit with dual feedback loop [J]. Electronics Letters,2011,47(24):1311 – 1313.

DOI:10.16644/j.cnki.cn33-1094/tp.2015.05.009

移动IPv6切换技术研究

李 旭，秦 军，杨 昭

(南京邮电大学教育科学与技术学院，江苏 南京 210023)

摘 要：由移动IPv4技术发展而来的移动IPv6技术应用前景可观，但其存在诸多问题，如移动节点在不同网络间移动带来的网络切换问题，切换过程中地址重复检测的延迟问题等。基于层次化的移动IPv6，详细阐述了目前移动IPv6的几种切换技术，并对现有的几种切换技术在切换时延方面进行了比较，发现层次型快速切换技术有更小的切换时延和丢包率。

关键词：移动IPv6；切换技术；移动检测；重复地址检测；切换延迟

中图分类号：TN915.04 **文献标志码**：A **文章编号**：1006-8228(2015)05-26-03

Research on mobile IPv6 handoff

Li Xu, Qin Jun, Yang Zhao

(*College of Education Science and technology of Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210023, China*)

Abstract: The mobile IPv6 technology developed from the mobile IPv4 technology has a considerable application prospect, but there are many problems, such as the network switching problem when mobile nodes roam between the networks, the handoff delay problem because of duplicate address detection in switching process, and so on. In this paper, several existing mobile IPv6 handover technologies are elaborated and their handoff delay are compared, and the result found that the fast hierarchical handover technology has a lower handoff delay and packet loss rate.

Key words: mobile IPv6; handover technology; mobile detection; duplicate address detection; handover delay

0 引言

近几年网络技术快速发展，下一代网络(NGN)将是今后通信业务和互联网业务的核心。以IPv6作为内在组成部分的移动IPv6技术，对于下一代移动通信网络有着极其重要的影响。IPv6有大量的地址资源和其他先进的性能，使网络地址转换(NAT)通信模式向对等网络(P2P)模式转换，解决了IPv4存在的一些关键性问题。但是基本移动IPv6协议仍存在较多的问题需要解决，如安全、AAA(身份认证、授权机制、自动计费服务)、切换延迟、组播等。移动IPv6技术将是下一代网络的核心，因此，有必要进一步认识和深入研究移动IPv6技术。

1 移动IPv6(MIPv6)

在传统的IP网络上，当移动节点(MN)离开一个网段而连接到新的网段时，需要给移动节点配置不同的IP地址，否则它不能按传统的路由机制将数据包路由到移动节点现在的位置，从而导致通信终断[1]。为了保持MN在移动中会话的连续性，相关研究组织提出了移动IP网络。已有出版的移动IP：Mobile IPv4(MIPv4)和Mobile IPv6(MIPv6)。MIPv6是MIPv4的升级版，MIPv6借鉴了MIPv4的很多概念和想法，并提出一些创新机制，解决了MIPv4中出现的三角路由、安全问题等。在MIPv6中MN可以在任意网内随意漫游，当MN与一个子网断开时，该节点将自动连接到另一个网段，而无需要像传统Internet进行手动配置IP地址。

MIPv6为了实现通信在网络层移动过程中保持通信不断，其解决方案可以简单地归纳为三个方面[2-4]。

(1) 家乡地址，MN在家乡链路中所获得的IP地址，MN通过该IP地址与外部节点进行信息沟通，保证了对应用的移动透明。

(2) 转交地址，MN移动到外链路时，MN根据外链路的子网前缀信息和自身的链路层接口生成的一个IP地址，保证了现有路由模式下通信可达。

(3) 家乡地址与转交地址的映射，建立了应用所使用的网络层标识与网络层路由所使用的目的标识之间的关系。

在MIPv6网络中，MN从一个网络自动转接到另一个网络，并保持其网络连通性的过程叫切换[5]。当MN在家乡网中，MN与通信节点之间按照传统的路由技术进行通信。当MN移动到外地链路时，MN的家乡地址保持不变，并获得一个转交地址，MN把家乡地址与转交地址的映射告知家乡代理。通信节点与MN通信仍然使用MN的家乡地址，数据包仍然发送到MN的家乡网；家乡代理截获这些数据包后，根据已获得的映射

收稿日期：2015-2-13

作者简介：李旭(1995-)，男，云南昭通人，在校本科生，主要研究方向：计算机网络，移动互联网技术等。

关系通过隧道方式将其转发给MN的转交地址，MN则可以直接和通信节点进行通信[6-7]。工作原理如图1所示。

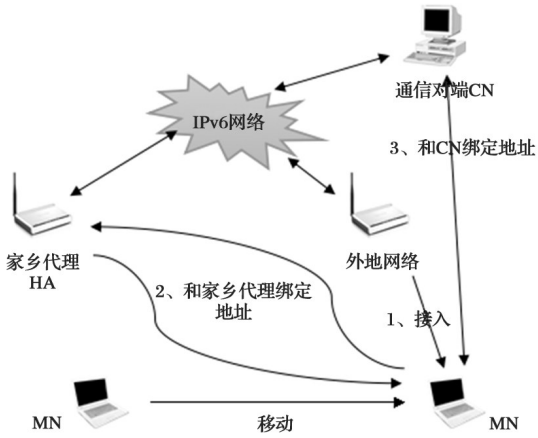


图1 工作原理

2 移动IPv6切换延迟

在MIPv6网络中，MN在不同网络间切换时先执行链路层切换后执行网络层切换，在这段期间MN既不能发送，也不能接收数据包，导致通信的终断，造成较大的切换延迟。切换延迟如图2所示。

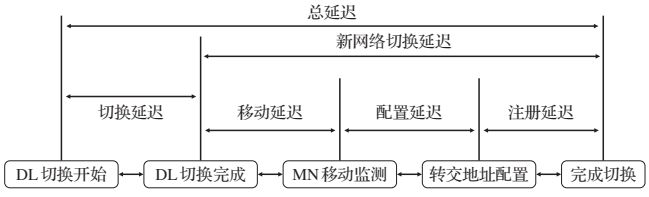


图2 切换延迟

由此可知，存在数据链路层切换时延T_{DL}，网络层移动检测时延T_{MD}，转交地址配置时延T_{COA}，重复地址检测时延T_{DAD}，绑定更新时延T_{BU}。为了改善MIPv6的切换性能，IETF提出了以下改进协议：MIPv6快速切换技术(FMIPv6)，MIPv6层次化切换技术(HMIPv6)和MIPv6层次型快速切换技术(F-HMIPv6)。

3 移动IPv6快速切换技术(FMIPv6)

FMIPv6采用链路层触发的方法预测切换的发生，将网络层切换的部分操作提到链路层切换之前，通过提前预测MN的移动位置，配置$NCoA$，进行DAD（重复地址检测）过程，加快了切换过程的完成[8]。切换过程如图3所示。

⑴ 移动节点由链路层触发机制预测到自己将要发生移动时，移动节点向前接入路由器(PAR)发送路由器请求代理消息。

⑵ PAR返回代理路由器通告消息，告知新接入路由器(NAR)的信息。

⑶ MN形成新转交地址($NCoA$)，并将其包在快速绑定更新(FBU)消息中发送给PAR。

⑷ PAR收到FBU消息后在$PCoA$和$NCoA$之间建立隧道。然后向NAR发发起切换消息(HI)，HI消息中包含了MN的$NCoA$。

⑸ NAR对$NCoA$进行DAD操作，检查$NCoA$是否有效。如果地址无效，NAR会重新给MN分配一个$NCoA$，并在切换确认消息(HA_{CK})中将结果返回给PAR。

⑹ PAR收到HA_{CK}后，向 MN和NAR返回FB_{ACK}消息，将发往$PCoA$ 的数据通过隧道送至NAR，NAR将数据包暂时缓存起来。

⑺ MN到达新的子网，向NAR发送快速邻居通告消息(FNA)，可从NAR接收到缓存的或新来的数据。

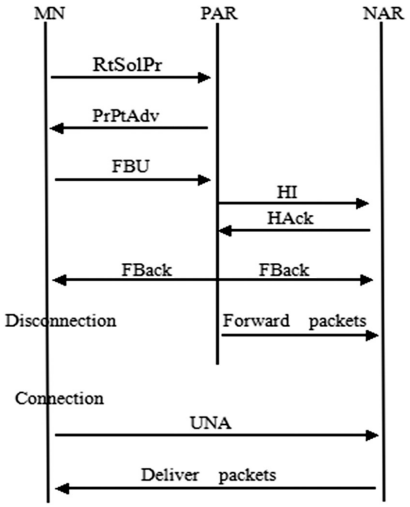


图3 切换过程

从上述分析可以看出，MN在连接到新的子网之前，已经获知新子网的信息并配置了经过DAD的$NCoA$，由此可知，预测式快速切换可以大大减少网络层移动检测和配置CoA的时间，减少了数据的丢包率。

4 移动IPv6层次化切换技术(HMIPv6)

无论是MIPv6还是FMIPv6，都存在切换时延较大和网络负荷过重等问题。于是IETF提出了层次化的MIPv6切换技术HMIPv6[9]。HMIPv6利用区域划分的思想，在逻辑上将网络划分成不同的域，每个域由一个称为“移动锚点”(MAP)的实体来管。一个MN在一个MAP域内有两个CoA，分别是$RCoA$和$LCoA$[10]。

当MN发生了域内移动时，MN通过RA报文配置新的$LCoA$，此时MN的$RCoA$对HA和CN仍然有效。当MN发生域间切换时，其步骤如下。①MN首先通过RA报文，获取AR的子网前缀和MAP的子网前缀，然后通过参数设置选择无状态的地址配置方式配置$LCoA$和$RCoA$。②MN向NMAP发送包含$RCoA$和$LCoA$域内绑定更新的LBU报文后，NMAP更新自己的缓存机制，更新MN的$RCoA$和$LCoA$的映射关系。③NMAP向MN发送LBA报文，表明注册成功，MN向HA发送BU报文，HA更新自己的绑定缓存记录。④MN向PMAP发送$PRCoA$和$NRCoA$的对应关系，PMAP和NMAP之间建立了隧道机制。⑤当CN向MN发送数据时，CN首先检查它的绑定缓存，检查MN的$RCoA$和HA的对应关系。然后CN更新自己的绑定缓存，记录MN的$RCoA$和HA的映射关系。此后，MN和CN将绕开HA直接进行通信。HMIPv6的网络流程如图4所示。

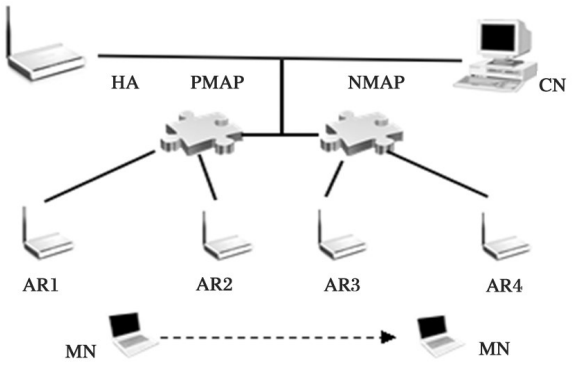


图4 HMIPv6的网络流程图

HMIPv6时延分析：MN在域内移动时，只需绑定新的AR和MAP，不需要向HA和CN发送BU报文，所以切换时延变得比原来小。MN在域间移动时，T_{BU}过程比原来要多几个步骤，所以域间移动的时延要比标准MIPv6切换时延大。

5 移动IPv6层次型快速切换技术(F–HMIPv6)

上面讲述到的两种切换技术，如果在较小局域内进行频繁移动时，可以使用HMIPv6来减少切换延时，而如果在层次MIPv6网络上应用FMIPv6，MIPv6的移动性将会得到极大的加强[11–12]。在F–HMIPv6中，建立MAP和NAR之间的快速切换的隧道，MN和MAP之间交换FMIPv6消息。F–HMIPv6切换过程如图5所示。

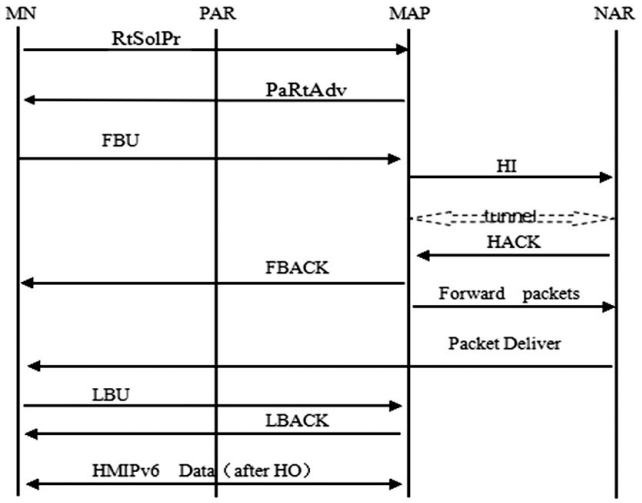


图5 F–HMIPv6切换过程

操作流程：由预期的切换，MN将发送路由器请求代理消息给MAP，MAP收到消息后会发送路由器通告消息回复MN。随后MN发送FBU给MAP。MAP在接收到FBU后会发送HI消息给NAR，确认切换后，MAP和NAR之间将建立一个双向隧道。MAP会根据PLC_oA和NLC_oA发送FBACK消息给MN。MAP通过双向隧道将发给MN的数据包转发给NAR并由NAR将数据包缓存起来。当MN移动到NAR的范围内，经过确认消息后，NAR会将刚才缓存起来的数据包通过NLC_oA转给移动后的MN。

F–HMIPv6时延分析：F–HMIPv6结合了FMIPv6和HMIPv6各自的优点，减少了T_{MD}、T_{COA}和T_{DAD}带来的时延。在微移动情景下，F–HMIPv6减少了T_{MD}、T_{COA}、T_{DAD}和T_{BU}带来的总时延，改善结果十分明显。

6 三种切换技术在时延上面的比较分析

在FMIPv6中，MN通过链路层触发机制减少了网络层移动检测和配置C_oA的时间，减少了数据的丢包率[7]。在HMIPv6中，当MN在域内移动时，只需绑定新的AR和MAP，当MN在域间移动时，T_{BU}会增大，故HMIPv6域内移动。在层次型快速切换技术(F–HMIPv6)中，FMIPv6主要减少了配置带来的时延，在微移动情况下，HMIPv6又减少了T_{BU}的时延。

由上述时延分析可知，三种切换技术都能有效的减少时延，其中层次型快速切换技术(F–HMIPv6)减少时延效果最好。

7 结束语

MIPv6切换性能已经成为阻碍MIPv6网络的实际应用和大规模商业化的最主要原因之一，因此降低切换时延有着重要意义。FMIPv6机制的提出降低了MIPv6的切换时延，当MN在域内移动时HMIPv6切换技术能减少信令负载，F–HMIPv6减少了移动、配置和重复检测带来的时延，使得F–HMIPv6在对实时性要求更高的商务活动中更趋于实用。虽然F–HMIPv6有效的减少了时延，但也实现不了无缝切换。同时影响MIPv6应用到实际通信中的因素还有安全性、服务质量等，这些问题都有待解决。

参考文献：

[1] 任孝金.IPv6技术与应用[M].清华大学出版社,2010.

[2] D. Johnson, C. Perkins, 1.Arkko. Mobility Support in IPv6[S].RFC 3775,June 2004.

[3] Optimistic Duplicate Address Detection (DAD) for IPv6[S]. RFC4429,April 2006.

[4] Cheng Y, Kao S, Chang F. Time–oriented care–of address for mobile IPv6 networks [C].2012 IEEE International Conference on Communication, Networks and Satellite,2012:74–78

[5] 蒋亮,郭健.下一代网络移动IPv6技术[M].机械工业出版社,2005.

[6] 张志群.移动IPv6切换技术研究[D].西安电子科技大学,2010.

[7] 林嘉燕,俞鹏伟.移动IPv6切换技术[J].计算机技术与发展,2008.10.

[8] 文雷飞.基于数据链路层移动IPv6快速切换方案的研究及实现[D].兰州大学,2007.

[9] R.Koodli. Fast handovers for mobile IPv6[S]. IETF RFC 5568, July 2009.

[10] H.Soliman. Hierarchical mobile IPv6 (HMIPv6) mobility management[S]. IETF RFC 5380, Oct 2008.

[11] 陈魏鑫,韩国栋,刘洪波.基于快速DAD的分层移动IPv6切换算法[J].通信学报,2008.1.

[12] 金源,李松年,张世永.基于F–HMIPv6的MAP自适应选择算法[J].计算机应用与软件,2007.9.