支持隐式人机交互的分布式视觉系统

王国建 陶霖密
（清华大学计算机科学与技术系，北京 100084） （普适计算教育部重点实验室，北京 100084）

摘 要 交互技术的发展推动着交互方式的转变并催生新的交互理论和交互范式。隐式人机交互是一种新型人机交互范式，其特点是用户无需直接操作计算机等设备，交互系统通过分布式传感器检测，分析用户动作，在线地获取当前的动态上下文，用于理解用户的意图，作为系统的隐含输入，产生与用户意图相适应的隐含输出来为用户服务。本文分析了隐式交互对计算系统的需求，提出了通用化的面向应用的服务共享模型，以用于构建分布式视觉系统。基于该系统实现的隐式交互实例表明，上述模型是通用的服务共享模型，能够实现多种隐式交互系统。

关键词 隐式交互 视觉系统 交互理论 上下文感知
中图法分类号：TP301.6 **文献标志码**：A **文章编号**：1006-8961(2010)08-1133-06

Distributed Vision System for Implicit Human Computer Interaction

WANG Guojian, TAO Linmi
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
(*Key Laboratory of Pervasive Computing, Ministry of Education, Beijing* 100084)

Abstract The development of interaction technologies is advancing new interaction models and theories. Implicit human computer interaction is a new interaction paradigm, which models a new interaction pattern that people is served without manipulating the interaction devices. The interaction system serves to people based on the understanding of their intention by analyzing the behavior and being aware of context in real time. Based on the analysis of the requirements of implicit interaction, this paper proposed a generalized application oriented service share model (AOSSM) for constructing distributed vision system to support implicit interactions. Two implicit interaction systems on group interaction and ambient kitchen are implemented, which demonstrated the effectiveness and the generalization of AOSSM.

Keywords implicit interaction, vision system, interaction theory, context aware

0 引 言

隐式人机交互是以人为本计算模式下的一种新型人机交互范式，是和谐自然交互的一个重要部分，其特点是用户无需直接操作计算系统，系统通过分布式传感器检测用户动作，在线获取当前的动态上下文，用于理解用户的意图，作为系统的隐含输入，产生与环境相适应的隐含输出来为用户服务。隐式交互的关键是系统要克服从低层传感器数据得到关于用户意图的高层理解之间的语义鸿沟，其中富有挑战性且具有普适意义的关键问题是在线动态上下文模型与感知，并以当前的动态上下文为桥梁，沟通语义鸿沟，而使得系统能逐步获得各层次的语义，逼近对用户意图的正确理解。实现隐式人机交互需要一定的软硬件环境，其中最重要的部分是分布式视觉系统，用于获取环境信息和用户的动作、表情等，并给出适当的反馈。隐式交互在会议系统、老年看护等各方面有着非常广泛的应用。

分布式视觉系统的研究和开发自本世纪初以

基金项目：国家自然科学基金项目(60873266,90820304)
收稿日期：2010-03-29；**改回日期**：2010-04-14

第一作者简介：王国建(1972—)，男。现为清华大学计算机科学与技术系计算机应用专业博士研究生。主要研究方向为计算机视觉、人机交互。E-mail：wgj06@mails.thu.edu.cn

来,一直是计算机视觉研究的一个热点,自 1999 年召开首届国际计算机视觉系统大会(ICVS)[1]以来,至今已经召开了 7 届,其中 ICVS2007 的主题是:“真实世界中的视觉系统”,ICVS2008 的主题是:“认知视觉系统”,直接面向现实世界中人的行为理解和自然交互。目前,许多研究机构研发了用于构建智能环境或自然交互的分布式视觉系统,如加拿大国家研究委员会提出的无缝消息系统[2]、微软公司提出的 Easy living 系统[3]等,MIT 的媒体实验室、CMU 的 AMI 实验室以及 Georgia Tech 等著名的大学还提出了感知家居等[4-6]研究项目。

视觉传感器具有信息丰富,对人体非侵入等特点,和主动传感器结合,可以在时空上对场景进行全面连续记录,为获取人的行为意图提供了可靠的基础。但视觉信息量大,处理复杂,占用较多的计算资源,对一些物理条件如光照、遮挡等变化比较敏感等因素,都影响传感器数据的实时处理和高层语义理解。因此如何基于分布式视觉系统真正实现隐式交互,还面临诸多挑战。目前许多大学和研究机构都对此给予了很多的关注,其研究内容涵盖了计算、认知和社会系统以及物理设备环境等[7]。例如,欧盟的 CANDELA 项目[8-9]考虑了视频和消息数据的传输,以及多机协作处理的问题,比较适合快速地构建基于视觉的原型系统;加州大学提出的 DIVA(distributed interactive video array)[10-11]系统中,设置了“Event Inferencing Modules”和“Infrastructure Modules”,提供给用户一系列虚拟传感器,可以用来配置和操作真正的传感器。这些系统都实现了部分的分布式视觉系统的功能,但是还缺乏从支持隐式交互的角度进行深入研究。

1 分布式视觉系统

近年来,视觉系统已经由原来的以记录面向室内或室外环境中的信息为主,向以人为中心的计算转变,即通过对获得的多路视觉信息进行处理、分析和理解,最终获取人的意图,从而向人提供主动服务。由于隐式交互本身所具有的特点,视觉系统也需要面对新的计算模式提出的一些基本要求。

1.1 具有交叉视域的多摄像机配置

隐式交互的中心是对人行为的理解,这就要求系统首先能对人的行为进行全程的跟踪和记录。用户在一定的环境中自由活动,单摄像机视域有限,对用户观察的角度单一,难以覆盖整个区域,遮挡问题更是难以独自解决,因此基于单摄像机的视觉系统无法胜任以人为中心的计算。系统需要分布式地配置多摄像机,而且对于重要区域摄像机之间应该有交叉视域以对人的动作进行不间断地跟踪和多角度地观察,为正确分析和理解人的行为提供客观可能。

1.2 实时的信息处理和理解

在隐式交互中,计算系统的主要目的是给用户提供主动的服务,要求在合适的时间向用户提供合适的服务,这是以系统能对信息进行实时处理为基础的。但是,在分布式视觉系统中,巨量的视频信息流、复杂的中间处理结果,以及基于概率图模型并融合各种信息的高层推理等,对信息系统来说,实时处理是一项极具挑战性的任务。进行分布式处理是基本的方向,可以保持系统的可扩展性并可持续提高其处理能力,这同时也需要系统充分考虑内部各模块的通信协作问题。

1.3 上下文处理的支持

隐式交互的基础是理解人的意图。在人与人及人与其周围事物之间的交互中,人的行为,如肢体动作、话语等在不同的环境(氛围)传达不同的意图。换言之,理解人的意图离不开当时的上下文信息,即动态上下文。这就要求系统在设计上必须考虑增加处理上下文的能力。另外,在一个视觉系统里,有限的计算资源和繁重的处理任务之间将会一直存在矛盾,如果具有上下文处理能力,则可以利用上下文导引数据的处理,突出处理的重点,既能解决计算资源和处理任务之间的矛盾,又可以加快处理速度。

1.4 历史信息管理

人的交互习惯,行为模式,以及个人偏好等都是重要的上下文信息,可以对当前及以后的交互行为进行预测。另外,群体交互等应用也要求系统具有记录及查询历史信息的功能等。因此,隐式交互要求信息处理系统具有完备的历史信息管理功能,主要包括数据存储、访问、回溯等功能。

综上所述,隐式交互的实质是在动态环境下理解人的意图,对视觉系统而言,就是分布式多模态信息获取、分布式信息处理,以及动态上下文的获取、管理和基于当前上下文的推理。最终目标是一个具有场景分析识别、行为理解能力的基于多摄像机的分布式视觉系统。

2　面向应用的服务共享模型(AOSSM)

在一个视觉系统中,一般包含检测、跟踪、识别等多个视觉算法模块,有些算法本身还具有不稳定、运算复杂等特点,因此视觉系统的稳定性和可靠性成为其最终能否应用于实际的一个关键问题,另外,由于上下文环境的变化及用户需求的变更,上层应用随之经常产生对于计算服务不同的需求,这需要系统能够动态地围绕上层应用提供满足其需求的计算服务。因此,使系统更好地满足上层应用对计算服务的动态需求,是实现基于分布式视觉系统的隐式交互的关键因素。目前很多研究机构提出的视觉系统不能真正用于隐式人机交互,也主要是因为其运行不能动态适应环境条件的变化,从而使其只能应用于条件单一的场合如工业检测等。对此,在总结现有视觉系统体系结构的基础上,提出一种通用的面向应用的服务共享模型(AOSSM),其设计思路是对包括视觉计算在内的各种计算服务,例如数据采集、视频处理、语义推理等进行抽象分层,使服务资源标准化,目标是为上层应用提供更加灵活的服务视图,实现分布式服务资源共享和按需分配使用。

如图 1 所示为该模型的结构示意图,总体上该模型分 3 层,分别由 3 个空间完成相应功能。AOSSM 的底层是系统服务空间,负责管理视觉系统能够提供的实际服务,包括传感器数据处理,如采集、传输、资源分配等,但该层的重点放在视觉算法功能集成上,其与中间层的接口主要描述信息处理(主要是视觉处理)的实际状态及计算的属性等,向上层提供如位置、用户动作、环境配置等较低级的语义信息,并可按照上层的需求对本身的计算进行调整(或重新组合);AOSSM 的中间层是虚拟服务空间,主要负责对系统服务空间提供的信息进行整合处理并进行抽象化,屏蔽实际服务的接口,向上层提供统一的标准服务接口,以响应上层具有更高层语义的服务需求,根据上层应用的需求变化,其提供的接口成为沟通底层和最上层应用的桥梁,能够根据服务选择策略计算出满足应用需求的最佳服务组合并向底层发出计算变更的要求;AOSSM 的高层是应用空间,主要负责处理用户定制的服务需求(比如在会议场景中定制监控“4 人小型会议”的服务),并依此与虚拟服务空间交互向中间层提出需求。AOSSM 是一种分布式计算模型,其设计的主要功能是:

1)个性化服务　通过系统服务资源的组织管理为上层应用提供个性化的计算环境;

2)透明访问　向上层应用提供透明的访问服务是本模型追求的目标之一,上层应用提交相应的服务需求,便可以访问所需的服务(组合),而不必关心这些服务资源所在位置等细节;

3)服务资源的按需使用　根据应用需求和当前上下文为上层应用动态地选择所需的服务资源;

4)通用性　可以支持多种视频、音频等媒体数据,支持多种视觉算法服务。

AOSSM 中的应用空间是面向应用的,它将采用更高级、更容易为人理解的方式来表示和描述应用

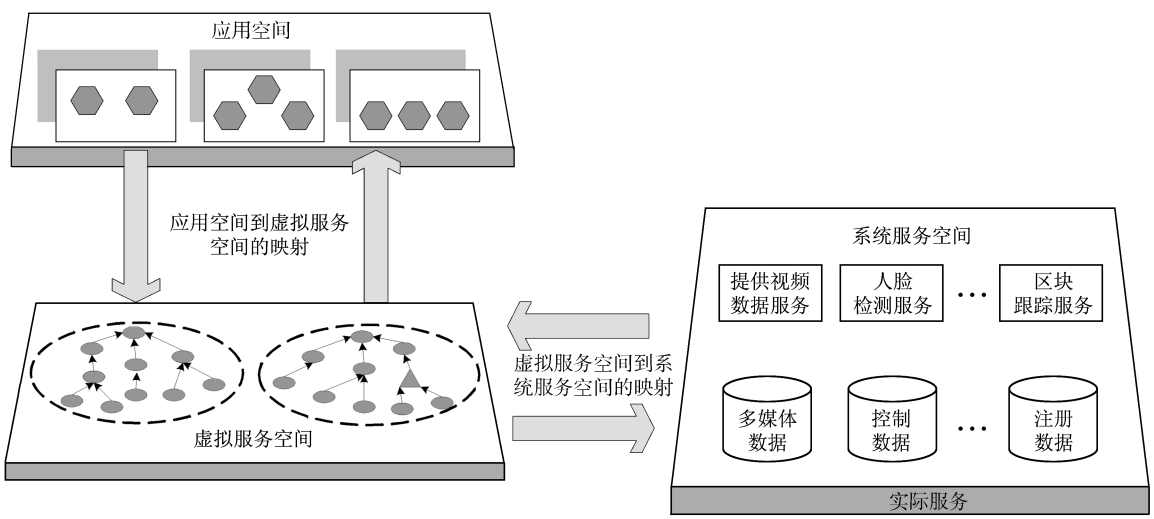


图 1　面向应用的服务共享模型 AOSSM 结构示意图

Fig. 1　Architecture of the application oriented service share model

资源,同时将根据应用的个性化需求重构当前所需的计算环境,这种重构是动态的。抽象服务空间是整个模型的关键,一方面它可以屏蔽底层不同提供者所提供的服务资源在语义描述上的差异,为上层应用提供标准化的语义支持;另一方面它将负责根据应用需求实现服务资源的选择,支持应用接口到应用实例的动态绑定,以实现对服务资源的智能管理。系统服务空间主要负责对有效服务进行注册和监控及其实际计算时的参数设定等。

有了这个概念模型后,为实现对视觉系统各种资源包括网络传输、视觉计算等进行有效的组织管理提供了参考依据,使得隐式交互的各种应用能够基于分布式视觉系统来实现。

3 基于 AOSSM 的分布式视觉系统

不同的分布式视觉系统虽然其应用的具体领域有所不同,但其一般都涉及到多路摄像机的数据采集、分布式处理以及信息融合等任务,由于在视觉系统里对视频数据的处理是最重要的并且也是最耗费计算资源的部分,对此从系统结构上需要设置多台计算设备,以随时应对新增加的需求,从而适应不同的隐式交互环境和需求。

基于上述隐式交互的实际需求和服务共享模型 AOSSM,提出了分布式视觉信息系统(DiVIP)体系结构。参考服务共享模型的定义,该体系结构对信息处理区分为 3 个层次:最里层主要处理传感器数据,中间层是各种计算模块和网络传输模块,最外层是各种具体应用模块的集成,可以调用组合中间层的计算服务。

如图 2 所示,DiVIP 是一种主要用于分布式视觉信息处理的多服务器平台[12],其基本思想是将视频以及其他传感器数据的采集、传输、存储等公共的部分抽取出来,封装成标准化的统一服务,同时将一些通用的视觉计算如检测、识别、跟踪等也封装起来以向更高层的应用进行语义分析提供服务,这样的平台实现了对传感器等底层数据处理的屏蔽,一方面使系统具有良好的通用性和可扩展性;另一方面高层应用的分析推理与底层传感器数据处理相互隔离,可以大大简化上层分析推理模块的开发并提高系统的可靠性。

根据本文提出的服务共享模型和 DiVIP 体系结构,具体设计了一个分布式视觉系统平台,目前主要

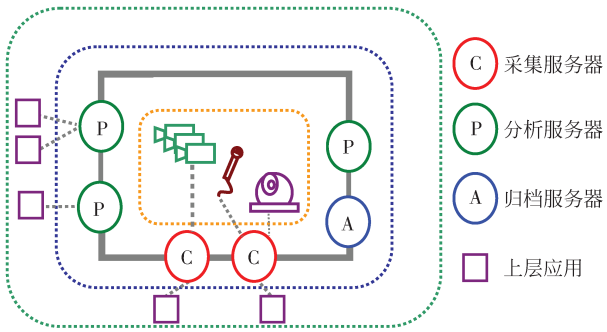


图 2 DiVIP 系统体系结构

Fig. 2 Architecture of the DiVIP system

实现了 AOSSM 模型的底层和中间层功能,如图 3 所示,主要由 5 个部分组成:1) Core Center,主要负责维护整个系统的配置,如系统有多少模块、各模块之间的关系等。2) Core,主要负责管理维护信息(数据)队列,包括源数据(传感器数据)和处理结果队列等,该部分完成了采集服务器的功能,同时为其他处理模块提供信息共享的平台,是 DiVIP 的核心。3) Module Runner,主要负责管理信息处理模块特别是视觉处理模块的加载运行,加载哪些模块以及怎样加载运行是该部分的关键功能。4) Module Dll ,实际的算法库,包括视觉算法以及完成其他功能如推理等的算法,可以针对具体的要求被 Module Runner 加载到系统以完成相应的功能。5) Plugin DLL,主要是用来帮助 Module DLL 获得更准确更特定的结果,例如某算法的不同参数、形状的不同描述等,增加了系统的可扩展性和灵活性。

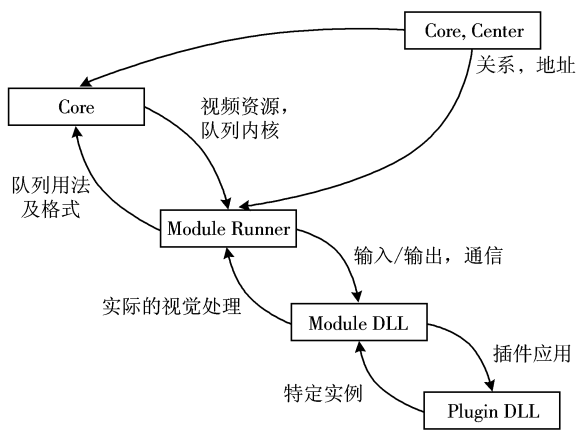


图 3 系统主要组成部分

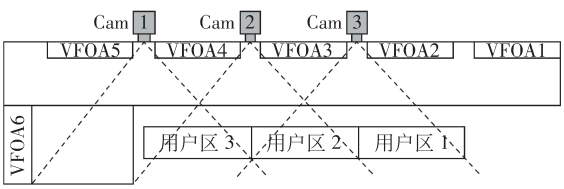
Fig. 3 The main components of system

系统平台的这几个组成部分中,Module Runner 主要用于完成 AOSSM 模型中间层的功能,实现计

算服务资源的选择策略。其他部分完成模型底层的功能,包括信息(视频)采集、数据(视觉)处理、网络传输等。各组成部分之间具有层次上的透明特征,即各层次之间具有很小的依赖性,这保证了功能模块具有很好的扩展性,使系统能够灵活地组织多个功能模块特别是通用化的视觉处理模块去理解场景,为隐式人机交互提供基础性服务。

目前我们提出的系统平台在人机隐式交互上已经实现了多种应用。与 New Castle 大学合作研究的智能厨房项目,最终目标是使系统根据烹饪的具体过程自动记录,人们通过视频回放分享烹饪体验。这就要求多个摄像机协作,在合适的时间运用合适的摄像机记录烹饪活动,比如在切菜时主要记录手部切菜的动作等。基于该系统平台提供的各种计算服务,结合主动传感器的运用,系统在对人进行全程跟踪的基础上,能够更加准确地检测事件和理解场景,也就可以自动记录最合适的烹饪视频片段。为此,对其原来只运用主动传感器的实验环境进行了改进,加装了监视用户正面烹饪操作的 3 个摄像头和用于全程跟踪的两个摄像头,图 4 为厨房工作台部分 3 个正面摄像头工作的示意图。目前初步实现的功能是利用分布式视觉系统平台提供的视觉处理基础服务,对用户进行全程跟踪,在此基础上重点研究头部姿态信息,得到用户关注焦点,以在用户面前自动显示或展开其关注内容的方式为其提供主动服务[13],实现人机隐式交互。

(a) 4 个显示区域,以及一些用于实验的标记



(b) 摄像机及兴趣区域的布局设置

(c) 当用户站在区域 2 时 3 个摄机分别捕捉到的图像

图 4　智能厨房有关实验环境

Fig. 4　Environment setup in the ambient kitchen

在隐式群体交互方面,如图 5 所示,是一个多人小型会议场景,利用分布式视觉系统,结合会议不同状态的上下文信息,可以对多人交互的情形进行处理分析,由于系统各层之间具有信息双向交流能力,能够通过自底向上和自顶向下信息处理的融合,不断得到更加准确的当前上下文[14],进而为用户提供更多主动式的服务,比如会议进程的辅助控制、视频数据的标注存储等。

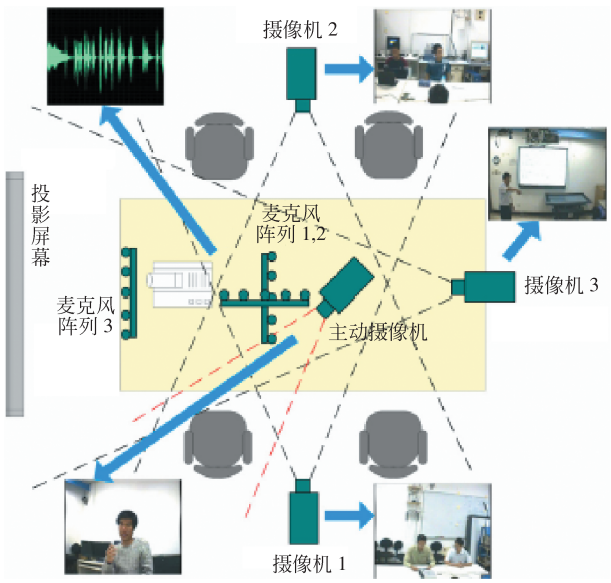


图 5　智能会议室环境配置

Fig. 5　Environment setup in the intelligent meeting room

4　结　论

视觉系统正在向隐式交互,在一定的场景中从人的行为理解人的意图并提供主动服务的方向转变,我们提出了支持隐式交互的服务共享模型,并基于该模型构建了视觉系统,实现了多种隐式交互。但是,这些隐式交互的实现仅是对隐式交互研究的开始,构建分布式视觉系统以实现在一定的场景中理解人的意图,仍面临着诸多挑战。

1)视觉信息处理的不确定性　传感器信息总是易受干扰,视觉计算本身也存在不可靠性,因此,如何构建分布式视觉或多传感器系统来降低这种不确定性,是今后分布式传感器系统所面临的一个重要课题。

2)动态上下文模型与觉察　上下文信息在视觉系统中对于行为理解具有重要作用,既是上层推

理的依据,又可以对底层数据处理提供导引,这已经成为共识。如何构建动态上下文模型,实现上下文觉察正在成为人的意图理解的关键问题之一。

3)隐式交互本体与系统　本体论的概念来源于哲学领域,是一种有效地表现概念层次结构和语义的模型。在隐式交互系统中,本体用于表达人的常识,是人的行为理解必不可少的一部分。如何表达人的常识并建立隐式交互本体,是隐式交互研究的根本问题。

近年来,视觉计算研究逐渐服务于人本计算,未来的视觉系统将是一个基于多摄像机且融合其他传感器的分布式系统,这种系统具备动态上下文觉察功能,能够基于交互本体和上下文信息实现高层的推理。我们认为这需要一个通用的参考模型支持,为此提出了以应用为中心的服务共享概念模型,并提供了基于该模型实现的两种隐式交互系统。我们相信,隐式交互系统必将更好地融入到人们的日常生活中。

参考文献(References)

[1] Michael Ley. List of International Conference on Computer Vision System(ICVS)[EB/OL]. [2010-3-15] http://www.informatik.uni-trier.de/~ley/db/conf/icvs/index.html.

[2] Miao Zhenjiang, Roger Impey, Yuan Baozong. Seamless messaging in a multimedia network environment [C]// Proceedings of the 6th International Conference on Signal Processing(ICSP'02). Beijing, China: Publishing House of Electronics Industry, 2002: 1031-1034.

[3] Steve Shafer, John Krumm, Barry Brumitt, et al. The New EasyLiving Project at Microsoft Research [EB/OL]. [2010-5-27]. http://research.microsoft.com/en-us/um/people/jckrumm/Publications%201998/smart%20spaces%20workshop.pdf.

[4] Ambient Intelligence Group. The Introduction of Current and Old Projects[EB/OL]. [2009-9-15]. http://ambient.media.mit.edu/projects.php.

[5] Ambient Intelligence Lab. The Introduction of Current and Old Projects[EB/OL]. [2009-9-18]. http://www.cmu.edu/vis/projects.html.

[6] Aware Home Research Initiative. The Introduction of Research [EB/OL]. [2009-8-13]. http://awarehome.imtc.gatech.edu/.

[7] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, et al. Guest editors' introduction: human-centered computing toward a human revolution[J]. Computer, 2007, 40(5): 30-34.

[8] RGJ Wijnhoven, EGT Jaspers. Flexible surveillance system architecture for prototyping video content analysis algorithms [C]//Proceedings of the International Society for Optical Engineering. San Jose, CA, USA: SPIE, 2006, 6073: 247-255.

[9] Jaspers E, Wijnhoven R, Albers R, et al. CANDELA—Storage, analysis and retrieval of video content in distributed systems [G]//Detyniecki et al. Lecture Notes In Computer Science. Heidelberg: Springer, 2006: 112-127.

[10] Mohan M Trivedi, Tarak L Gandhi, Kohsia S Huang. Distributed interactive video arrays for event capture and enhanced situational awareness[J]. IEEE Intelligent Systems, 2005, 20(5): 58-66.

[11] Trivedi M M, Huang K S, Mikic I. Dynamic context capture and distributed video arrays for intelligent spaces [J]. IEEE Transactions on Systems, Man And Cybernetics: Part A, 2005, 35(1): 145-163.

[12] Wang Yao, Tao Linmi, Liu Qiang, et al. A flexible multi-server platform for distributed video information processing [C]//The 5th International Conference on Computer Vision Systems. Bielefeld Germany: Bielefeld University, 2007: 21-24.

[13] Dong Ligeng, Di Huijun, Tao Linmi, et al. Visual focus of attention recognition in the ambient kitchen[C]//Proceedings of the 9th Asian Conference on Computer Vision. Berlin/Heidelberg: Springer, 2010: 548-559.

[14] Dai Peng, Tao Linmi, Xu Guangyou. Event based dynamic context model for group interaction[J]. The Official Journal of the Biomedical Fuzzy Systems Association, 2008, 13(2): 67-74.