The Emergence of Visual Crowdsensing: Challenges and Opportunities

Bin Guo, *Senior Member, IEEE,* Qi Han, *Member, IEEE,* Huihui Chen, Longfei Shangguan, *Member, IEEE,* Zimu Zhou, *Member, IEEE,* Zhiwen Yu, *Senior Member, IEEE*

***Abstract*—Visual Crowdsensing (VCS), which leverages built-in cameras of smart devices to attain informative and comprehensive sensing of interesting targets, has become a predominant sensing paradigm of mobile crowdsensing (MCS). Compared to MCS tasks using other sensing modalities, VCS faces numerous unique issues, such as multi-dimensional coverage needs, data redundancy identification and elimination, low-cost transmission, as well as high data processing cost. This paper characterizes the concepts, unique features, and novel application areas of VCS, and investigates its challenges and key techniques. A generic framework for VCS systems is then presented, followed by discussions about the future directions of crowdsourced picture transmission and the experimental setup in VCS system evaluation.**

***Index Terms*—Visual crowdsensing; mobile crowdsensing, object imagery; data selection; visual data understanding; crowd intelligence.**

I. Introduction

With the development of smartphone sensing, wearable computing, and mobile social networks, a new sensing paradigm called Mobile Crowd Sensing (MCS) [1, 2], which leverages the power of regular users for large-scale sensing, has become popular in recent years. Data collected onsite in the real world, combined with the support of the backend server where data fusion takes place, makes MCS a versatile platform that can often replace static sensing infrastructures.

MCS can make use of different modalities of sensing, e.g. numeric values (e.g., air quality [3], GPS coordinates [4]), audios, and pictures/videos. Among these modalities, visual crowdsensing (VCS) that uses built-in cameras of smart devices has become increasingly popular. VCS asks people to capture the details of interesting objects/views in the real world in the form of pictures or videos. It has attracted considerable attention recently due to the rich information that can be provided by images and videos. Previous projects, e.g. CreekWatch [5], GarbageWatch [6], PhotoNet [7], PhotoCity [8], WreckWatch [9], FlierMeet [10], and Mediascope [11], indicate that VCS is useful and in many cases superior to traditional visual sensing that relies on deployment of stationary cameras for monitoring.

Compared to other sensing modalities (e.g., numeric values, audios) in MCS, images/videos are more informative (e.g., rich objects captured), richer in associated contexts (e.g., shot size, shooting angles), larger in data item size, and more complex on data processing. Furthermore, VCS faces several unique issues, such as *multi-dimensional coverage needs*, *data redundancy identification and elimination*, *low-cost transmission*, and *high data processing cost*. Though VCS has been used in many applications, there has been no comprehensive investigation of this field. In our previous work [1], a systematic review of generic MCS concepts, applications, and research issues are presented. However, it did not characterize the unique features and challenges of VCS, the rich VCS applications and associated techniques are not investigated as well. To this end, this paper aims to provide a thorough review of the research issues and state-of-the-art techniques, and present our insights of VCS. In particular, we have made the following contributions.

(1) Characterizing the concepts and features of VCS, including its working process, data coverage and redundancy, crowd-object interaction, and crowd intelligence. A generic concept model is further presented.

(2) Reviewing existing VCS applications on object imagery and profiling, visual event sensing, disaster relief, localization, indoor navigation, and personal wellness.

(3) Investigating the challenges and key techniques of VCS including diversity-oriented task allocation, data selection and redundancy elimination, opportunistic visual data transmission, energy-efficient and reliable communication, image matching and processing, picture quality estimation, and visual data understanding.

(4) Presenting our efforts and the future trends of VCS, giving a generic framework for VCS systems, discussing the future direction on integrating with new communication techniques, using crowd intelligence for crowdsourced visual data understanding, and summarizing the experimental setup in VCS evaluation.

The remaining paper is organized as follows. In Section II and III, we characterize the unique features of VCS. Section IV

This work was partially supported by the National Natural Science Foundation of China (Nos. 61332005, 61373119, 61602230), and the National Basic Research Program of China 973 (No. 2015CB352400).

B. Guo, Huihui Chen, Zhiwen Yu is with Northwestern Polytechnical University, Xi'an, 710072, China (e-mail: guob@nwpu.edu.cn, chenhuihui.cn@gmail.com, zhiwenyu@nwpu.edu.cn).

Qi Han is with Colorado School of Mines, 1500 Illinois Street, Golden, CO 80401, USA (e-mail: qhan@mines.edu).

Longfei Shangguan is with 35 Olden Street, Princeton, NJ, 08540 (e-mail: longfeis@cs.princeton.edu).

Zimu Zhou is with ETH Zurich, ETZ G85, Gloriastrasse 35, CH-8092 Zurich, Switzerland (e-mail: zzhou@tik.ee.ethz.ch).

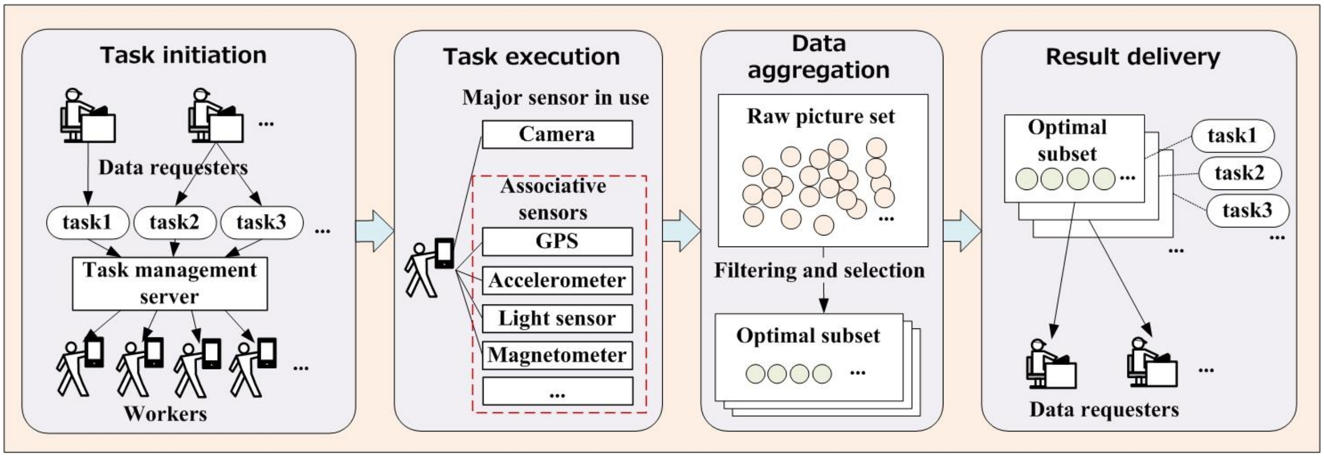


Fig. 1. VCS work flow.

classifies various novel applications enabled by VCS, followed by the challenges and key techniques discussed in Section V. Our insights and future research directions are discussed in Section VI. We conclude the paper in Section VII.

II. Visual Crowdsensing: An Overview

Before tackling the technical details, we first present an overview of the development of VCS, its relationship with mobile crowdsensing, and its generic working process.

A. The Development of VCS

In [12], Sheng et al. present the sensing as a service (S^2aaS) concept, which refers to smartphone-based sensing service provision via a cloud computing system. Mobile crowdsensing (MCS) presents a new sensing paradigm leveraging the power of mobile devices, which is a promising research area under the S^2aaS concept. According to [1], MCS is formally defined as: *the ability to acquire local knowledge through sensor-enhanced mobile devices and the possibility to share such knowledge within the social sphere, practitioners, healthcare providers, and utility providers*.

Visual Crowdsensing (VCS) is a specific form of MCS, which *tasks people to capture the details of interesting objects/views in the real world in the form of pictures or videos*. Following are several representative applications of VCS. SeeClickFix[1] allows people to report neighborhood issues (e.g., road collapse, public facility damages) to local government bodies in the forms of pictures or videos. PhotoCity [8] relies on the citizens to collaboratively acquire urban imagery (e.g., 3D street views) at a large scale. Movi [13] identifies highlights or interesting scenes from crowd-contributed videos to generate a visual summary of an event through collaborative sensing.

There have been other types (e.g., texts, audios) of multimedia applications of crowdsensing [14]. For instance, Sakaki et al. [15] investigate the real-time interaction of events (e.g., earthquakes) in Twitter and propose an algorithm for event detection by mining crowd-contributed tweets. NoiseTube [16] is an audio-based system for citizens to measure their personal exposure to noise in their daily lives and participate in the creation of noise maps. The difference between VCS and other crowdsourcing multimedia systems is that visual contents, i.e., pictures/videos, generally have high dimensional feature space and high transmission cost, resulting in significant burdens on computation and communication. In addition to spatial-temporal coverage needs, VCS tasks usually have more semantic coverage requirements (e.g., shooting angle and shot size), which raises new issues on task allocation and data quality measurement.

B. The Generic Work Flow of VCS Tasks

Data collection of a VCS app is usually conceptualized as a task in a traditional multi-task crowdsourcing platform, such as Amazon's Mechanical Turk (MTurk) [17] and Medusa [18]. A VCS task can be characterized by a generic four-stage process, as depicted in Fig. 1: *task initiation*, *task execution*, *data aggregation*, and *result delivery*. At the *task initiation* stage, *data requester*s define their tasks with different requirements and the *task management server* allocates the tasks to suitable workers or workers select their tasks by their own. At the *task execution* stage, *worker*s take pictures or videos according to task requirements and upload them to the backend server. Since the server receives pictures/videos uploaded by distributed workers intermittently, it is inevitable that there can be redundancy in pictures/videos and some user-contributed data items may be of low quality. As such, at the *data aggregation* stage, pictures/videos are grouped, filtered, and selected based on task specifications and data quality. In the *result delivery* stage, the data after preselection is made available to the data requesters upon task completion.

III. Characterizing Visual Crowdsensing

In this section, we introduce VCS as a special paradigm of MCS, emphasize on its unique characteristics compared with MCS using other sensing modalities, and formally define the concept models of VCS.

A. Data-Centric Crowdsourcing and Crowd-Object Interaction

[1] https://seeclickfix.com/

The ever increasing participants of crowdsourcing contribute large volume of data. Intelligently analyzing and processing crowdsourced data can maximize the usable information, thus paying back to the crowd.

There are several issues regarding crowdsourced visual data processing. First, the pictures/videos contributed by users are usually huge in quantity, while they vary in quality and reliability. Some people contribute accurate information (e.g., clear pictures) while others do not. Second, data from distributed 'human sensors' are often redundant, e.g., pictures taken nearby can be highly-duplicate. Third, crowd-contributed pictures/videos often contain rich associative information, such as geo-tags and picture-shooting contexts. These features make it challenging to analyze and understand crowd-contributed visual data. Existing methods are mostly based on the data itself. Analyzing the content of huge volume of data is usually computationally intensive and thus works poorly in many cases.

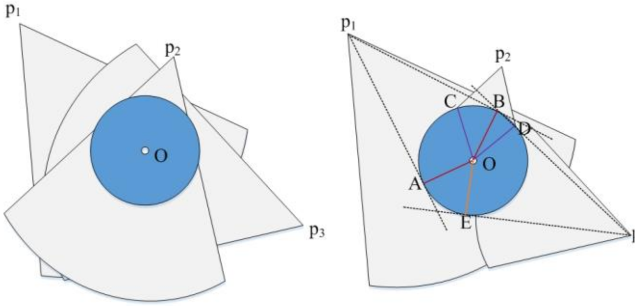
Generally speaking, VCS tasks are about *crowd-object interaction*, where people generate data about sensing objects in the real world. An in depth analysis of VCS (see Fig. 2) reveals three layers of information, including content, interaction context, and community context.

- *Content* refers to user-contributed pictures/videos.
- *Interaction context.* It refers to the relationship between human and data, i.e., how the data is contributed by human workers, such as time/space, interaction patterns.
- *Community context*. For a selected crowdsourced data set, there will be an associated community that participates in data contribution. Community contexts refer to the information regarding the community and its members, such as individual profiles or interests, social relationships, interaction dynamics.

Here, we term the interaction and community contexts as *crowd intelligence*. Crowd intelligence refers to aggregated human intelligence, which is formally defined as: *the context information generated during human-object interaction process or the associative information about the community and its members who contribute data*. In other words, crowd intelligence refers to the associative "information" (about the crowdsourcing task data) that can be obtained from the crowd-object interaction process and the relevant contributors. Later we will study how to measure and use them to facilitate crowdsensed visual data understanding.

B. Task Coverage and Data Redundancy

A VCS task may need the front, side, or back view of an object. Only knowing the location of the object is not sufficient, it is better to view it from multiple angles to obtain omnibus information. This is quite different from traditional sensor coverage, where the relative angle of sensors and targets does not matter because a target is considered covered as long as it is inside the sensing range of a sensor.



(a) point coverage (b) aspect coverage

Fig. 3. (a) For point coverage, the PoI *O* can be covered by photos *p1*, *p2* or *p3*. (b) For aspect coverage, aspect coved by *p1* is assessed by degree of ∠AoB. ∠CoD is for *p2* and ∠DoE is for *p3*.

Definition of Task Coverage. According to different task needs, coverage may vary in meanings. It is defined at the semantic level, using the constraints such as location, shooting angle, and shot size. We define the coverage in VCS at both *macro* and *micro* levels.

- ***Macro coverage***. It refers to the coverage of Point of Interests (PoIs) [19] defined by tasks. As shown in Fig. 3 (a), three pictures p_1, p_2 and p_3 all have the same macro coverage to PoI *O*. Therefore, if we want to have the information of the PoI at the macro coverage level, we can choose any one of them to complete the task. The situation is similar when we change pictures to video clips.
- ***Micro coverage***. It refers to multi-dimensional aspects of a PoI. As shown in Fig. 3 (b), three pictures p_1, p_2 and p_3 have different aspect coverage to the object at point *O* from different directions. If the task only requires two pictures, then $\{p_1, p_3\}$ will have the largest micro coverage. However, if the task requires 360-degree coverage, then

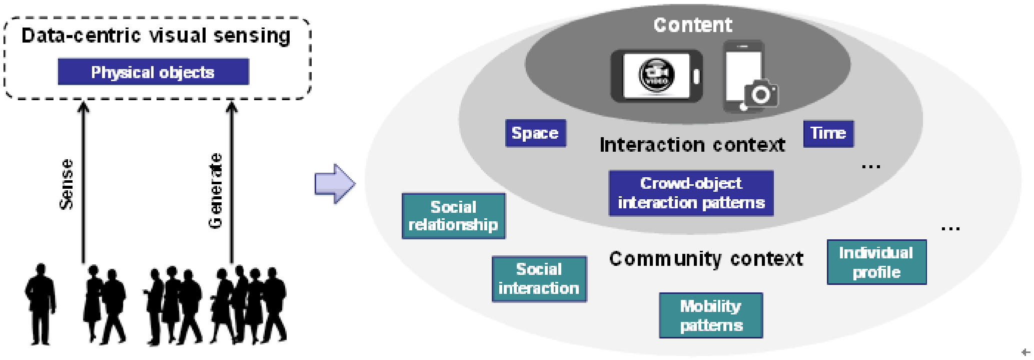


Fig. 2. Human and data-centric crowdsourcing: a deep insight.

{p_1, p_2, p_3} are all valuable and a picture that can cover ∠AoE is still needed.

The macro and micro coverage is similar to the *point* and *aspect* coverage concepts proposed by Wu et al. [20] and Wang et al. [21].

Definition of Data Redundancy. With coverage definition, there can be several pictures that cover the same point or aspect, resulting in data redundancy. The notion of redundancy is subjective, and more importantly, ultimately depends upon the intended use of the data or the definition of "coverage" in VCS tasks. For example, suppose that we have two pictures of the same building, if the goal is to detect the building, they are redundant; but if we want to provide a panoramic view of the building, they are not redundant. Furthermore, suppose that we have two video clips about an event from different locations. If the goal is only to detect the happening of the event, they are redundant; but if we want to characterize the event from different shooting angles, they are not redundant.

According to different task requirements, redundancy can have distinct meanings and can be roughly categorized into the following two types.

- ***Content-redundancy (ConR)***. This refers to the visual similarity among pictures or video frames based on visual features such as SIFT [22], color histogram [23].
- ***Semantic-redundancy (SemR)***. The similarity is defined at the semantic or contextual level, using features such as location [7, 24] or shooting angle [25]. For example, different buildings may look alike in pictures, but if their locations are different, there is no *SemR* because they carry distinct information.

C. VCS Concept Modeling

A VCS system is built on three key concepts, namely *task*, *user*, and *data*. We build a triple concept graph to characterize their underpinnings and relations in Fig. 4.

Task model. We propose a generic task model to characterize VCS tasks: *Task*=<*time*, *PoIs*, *w_num*, *c_set*>. Here, *time* is a valid period for performing the task, including the start time and the end time; *PoIs* refer to the target sensing areas. *w_num* refers to the number of workers needed for the task. *c_set* is the task-dependent *constraint set*. There are several often-used constraints. For example, *cg* is a geographical distance threshold, and data sensed within the range of *cg* could be semantically redundant; *ct* refers to the data sampling interval, and the data within the same interval can be considered redundant. There are constraints specific to pictures/videos, e.g., *ca* – the minimum orientation discrepancy of pictures/videos of the same target. Incentives are also important for a VCS task, and the task requester can state his/her budget for the task.

User model. The VCS tasks are conducted by participants, and thus we have the user model to characterize the participants. One relates to user profile, such as user name, age, profession, skills, interests, and preferences. The other refers to various user contexts, such as spatio-temporal contexts, mobility patterns, and social relations. The user model helps recruit appropriate workers to perform VCS tasks. It is also important for supporting user cooperation.

Data model. Each picture item *p* submitted is modeled as *p*=<*wid*, *cont*, *t*, *l*, *context_s*>. Here, *wid* refers to the worker *id* of the contributor; *cont* refers to the visual content; *t* and *l* denote *when* and *where* the data is obtained; *context_s* represents optional contexts of the picture/video. There are six often used contexts.

- The *shooting angle* of a picture or video, represented in the form: <*azimuth, pitch, roll*>, which can be obtained from accelerometer and magnetic field sensor readings [26]. It is a vector from the camera and vertical to the image plane.
- The *ambient light level* recorded by the light sensor.
- The *accelerometer readings* during photographing.
- The *depth-of-field* refers to the distance between the target and the camera, which is determined by four parameters: focal length, focus distance, lens aperture, and circle of confusion.
- *Field-of-view* refers to how wide the camera can see.
- *Effective range* of the camera, beyond which people can hardly identify anything in the picture.

When aggregated by the order of sensing time in the backend server, the data items form a data stream *P*. More specifically, it consists of a sequence of data items p_1,..., p_m,... arriving at timestamps t_1, ..., t_m, ...

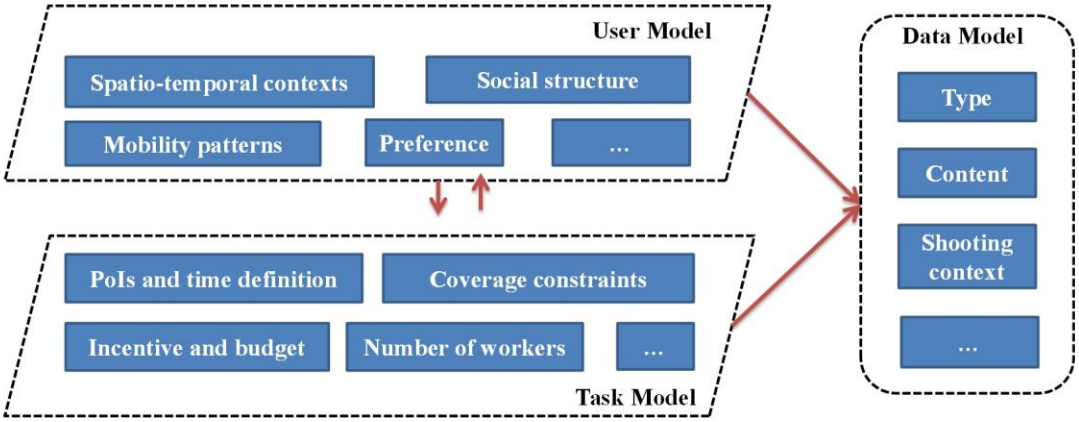


Fig. 4. The VCS concept graph.

In VCS, a picture stream is generated when pictures are being uploaded intermittently to the backend server by participants. Pictures contributed later in the stream may be semantically or visually similar to previous ones, resulting in data redundancy. Since a picture has heterogeneous features (e.g., *cont*, *t*, *l*, and *context_s* in the data model), we use a Boolean function $\mathcal{D}$ in Eq. (1) to measure the degree of duplication of two pictures p_i and p_j.

$$\mathcal{D}(p_i, p_j) = \wedge_{f=1,\ldots,k,\ldots d}\,\mathcal{H}(p_{i,k}, p_{j,k}) \qquad (1)$$

where f denotes the feature set and $p_{i,k}$ refers to the k-th feature value of picture p_i. The Boolean function $\mathcal{H}$ calculates whether two sub-items are similar. The calculation method of $\mathcal{H}$ can vary for different feature k. For instance, if k denotes locations, then $\mathcal{H}$ is a method (e.g., Euclidean distance) to determine whether two locations are close enough to take similar pictures.

When aggregating videos about a sensing target (e.g., a social event) capturing from different shooting angles or distances, data summarization or mashup is often needed. This will help choose views/frames from different videos to form a comprehensive picture about the sensing target. We discuss the details in the next section.

IV. Applications

To assist in identifying the needs of future VCS, we have developed a taxonomy of potential and existing application classes. The first division relates to the objects being sensed: stationary objects vs. dynamic events, as presented in Section III.A and III.B. The second division of applications relates to the purpose of crowdsourced visual data: disaster relief, indoor localization, indoor navigation, personal wellness, and urban sensing, as presented in the Sections III.C to III.G.

A. Stationary Object Imagery and Profiling

Object imagery uses crowdsensing to quickly make visual profiling of a physical object. Typical objects studied widely include floor plans, indoor/outdoor scenes, and so on.

Floor plan generation. The building floor plan is commonly used in architecture, showing the top-down view of the spatial relationships between rooms, spaces, and other physical features of a floor. It is vital for many indoor mobile applications, such as localization and navigation. CrowdMap [227] generates indoor floor plans by fusion of visual, inertial (e.g., gyroscope, accelerometer), as well as spatial information crowdsourced from people. Jigsaw [28] firstly uses visual and inertial data to infer the spatial relations, and then aggregates them to generate indoor floor plans.

Indoor scene reconstruction. Different from succinctly illustrating the floor plan of a building, scene reconstruction is to build visually appealing indoor interior views. It is useful for many applications such as virtual tours and indoor navigation. Existing outdoor street-view reconstruction techniques cannot be directly applied to indoor environments, while VCS introduces an effective and low-cost way to attain this. Sankar et al. [29] develop a smartphone app that lets users capture a panorama of indoor scenes. IndoorCrowd2D [30] is a VCS system that facilitates indoor scene reconstruction leveraging multi-dimensional sensing.

Outdoor scene reconstruction. There have also been studies on outdoor scene reconstruction for providing better location-based services. PhotoCity [8] leverages crowdsourced pictures for fine-grained building profiling. CrowdPan360 [31] uses crowd-sourced pictures to generate 360-degree panoramic maps when a user steps into an unfamiliar area. Kim et al. [32] develop a set of key-frame selection algorithms to automatically generate outdoor panorama using crowdsourced sensor-rich videos. RDB-SC [33] assigns visual spatial tasks (e.g., landmark profiling) to selected workers to enrich the spatial/temporal diversity of crowdsourced visual data.

B. Dynamic Event Sensing

With the prevalence of mobile Internet, more and more people record real-time events with their smartphones and instantaneously share pictures/videos through mobile social networks. This helps people quickly learn about the details of ongoing events, especially for those instant, ephemeral, and small-scale events, such as street performances, social events like parties, and meetings. InstantSense [34] leverages people's physical mobility and photographing to locate interesting events in real-time and further recount them with multi-grained and multi-facet visual summaries. Movi [13] enables smartphones to collaboratively sense their ambience and recognizes socially interesting events.

Each smartphone camera is able to capture only a range of restricted viewing angle and distance, which produces a rather monotonous video clip of an event. By spatial reasoning on the relative geometry of multiple video clips being captured from different angles and distances, FOCUS [35] can recognize shared contents and highlights of an event. With this, FOCUS supports real-time analysis and clustering of user-uploaded video clips about social events (e.g., a sport game). MoViMash [36] investigates how to combine crowdsourced event video clips to produce a more interesting and representative mashup of the event for sharing. A framework that supports smooth shot transitions to cover the performance from diverse perspectives is proposed. MoVieUp [37] is also a mobile video mashup system by learning film-editing rules from formal user studies. Frey and Antone [38] propose a cross-media tracking approach that can group crowdsourced mobile videos for event reconstruction.

C. Disaster Relief

Rapid disaster relief is important to save human lives and reduce property loss. Detailed and real-time information about the disaster area will help people make critical decisions on the assignment of manpower and supplies. The information, however, can be contributed by the rescuers, survivors and soldiers in the field by using their phones. One critical issue in disaster situations is that the network bandwidth is often limited. PhotoNet [7], CooperSense [39], and SmartPhoto [25] address this problem by only collecting and delivering a representative subset of pictures in crowd sensing, considering that a significant portion of the pictures may be redundant or irrelevant. CARE [40] designs a framework for better utilizing network resources in disaster-affected regions, which can detect the semantic similarity between crowdsourced photos. SmartEye [41] proposes a QoS-aware in-network deduplication scheme to attain efficient data sharing in disaster environments.

D.Indoor Localization

The current mainstream indoor localization technologies largely rely on RF signatures (WiFi access points [42], RFID [43]). Obtaining the signature map usually requires dedicated efforts to obtain fine-grained fingerprints. Alternative ways that have comparable performance or can provide complementary aid to existing techniques are being explored. The visual-based localization method, which leverages environmental physical features (e.g., logos of stores, paintings on the walls) as reference objects has been proved useful in real-world deployments. To alleviate the efforts in building visual reference object database, photo crowdsourcing is used. Xu et al. [44] employ structure from motion (SfM) to build the indoor 3D visual model from crowdsourced pictures, which is then used to solve the fingerprint ambiguity problem in indoor localization (two distinct locations may possess similar RF-fingerprints). Sextant [45] formulates visual reference object selection as a combinatorial optimization problem and proposes a heuristic algorithm based on iterative perturbation to enhance localization accuracy. CrowdSense@Place [46] identifies place categories (e.g., coffee shops, restaurants, meeting rooms) based on opportunistically collected photos and audio cues through smartphones. The place hints can be texts on signs or objects specific to an environment.

E. Indoor Navigation

Indoor navigation plays a significant role in complex indoor environments such as airports, shopping malls, and museums. A good indoor navigation system should supply the users with flexible navigation routes and user-friendly navigation instructions. Visual cues and image-based matching have been proved effective in indoor navigation, where VCS techniques have been found useful to lower the barriers to develop vision-based indoor navigation systems. iMoon [47] investigates the feasibility of utilizing crowdsourced data for building a smartphone-based indoor navigation system. It builds 3D models of indoor environment from crowdsourced 2D photos. With 3D models, it supports image-based localization and provides visual navigation instructions that show when and where to turn. Travi-Navi [48] is a vision-guided navigation system that enables a self-motivated user to easily bootstrap and deploy indoor navigation services. It records high quality images during the course of a guider's walk on the navigation paths, collects a rich set of sensor readings, and packs them into a navigation trace. The followers track the navigation trace, get prompt visual instructions and hints, and receive alerts when they deviate from the correct paths.

F. Personal Wellness and Health

TABLE I

A SUMMARY OF MAIN VCS APPLICATIONS AND TECHNIQUES USED

App type	Name	Modality	Technical Contributions
Object imagery and profiling	CrowdMap[27], Jigsaw[28]	Video, Image	Floor plan generation Efficient image matching
	Photocity[8]	Image	3D building modeling
	IndoorCrowd2D[30]	Image	Indoor scene construction, context-based quality estimation
	RDB-SC [33]	Video/Image	Reliable and diversity-oriented task allocation
Visual event sensing	Movi[13]	Video	Sensor-based data selection, Data understanding with crowd intelligence
	MoVieUp[37] MoViMash[36]	Video	Sensor-based data selection Content-based quality estimation
	InstantSense[34]	Image	Data selection, data understanding with crowd intelligence
	Frey and Antone [38]	Video	View matching and camera alignment
Disaster relief	PhotoNet[7], SmartPhoto[25] SmartEye[41]	Image	Redundancy elimination, selective data transmission
Localization	Xu et al. [44]	Image	Indoor 3D modeling
	CrowdSense@Place [46]	Image	Everyday object recognition, Context-based quality estimation
Indoor Navigation	iMoon[47]	Image	Indoor 3D modeling, fingerprint-based image matching
	Travi-Navi [48]	Image	Content- & context-based quality estimation
Public sensing	PetrolWatch [51]	Image	Data selection
	VizWiz[52]	Image	Task allocation
	DietSense[53], MT-Diet[54]	Image	Image tagging/classification, Content-based quality estimation
	PublicSense[55]	Image	Data analysis and visualization
	SignalGuru[56]	Video	Sensor-enhanced object detection
	SakuraSensor[59]	Video	Object detection, location-based image grouping
	GigaSight [60]	Video	Scalable infrastructure, privacy protection

People are misled into paying high prices to products due to the search costs on attaining price information. There have been several VCS studies that try to address this issue. For instance, LiveCompare [49] and MobiShop [50] are systems that allow for grocery bargain hunting through crowd photographing. They use barcode decoding and GPS/GSM localization to automate the detection of product identity and store location. PetrolWatch [51] uses mobile camera phones to collect, process and deliver pricing information from petrol stations to potential buyers. The main contribution is automatic billboard image captured from a moving car without user intervention. VizWiz [52] is a crowdsourcing app that allows blind people to post picture-based queries and receive answers from remote workers. Posted fliers on community bulletin boards advertise services, events, and other announcements, which serves as an important function for public information sharing in modern society. FlierMeet [10] is a crowdsensing system for cross-space flier information photographing and intelligent tagging. Dietary patterns are recognized as contributing factors to many chronic diseases. Logging dietary habits in the form of daily journals is thus important. DietSense [53] supports the use of mobile devices for automatic photographing of dietary choices and efficient tagging of the dietary images for querying and browsing. MT-Diet [54] is an automated diet monitoring app that combines infrared and color images to recognize food types and provides feedback to promote healthy eating habits.

G. Urban Sensing

MCS has become an important way to achieve large-scale urban sensing. The modern city encounters numerous municipal problems that may impact human daily life, such as noise disturbance, road collapse and public facility damage, such as street lamps and manhole covers. PublicSense [55] is an image-based crowdsensing system that allows citizens to give instant reports about public facilities. It has potential application areas such as public facility management, urban infrastructure maintenance, intelligent transportation services, and emergency situation monitoring. Similarly, SeeClickFix is a web-based service designed to help citizens report non-emergency issues in their neighborhood. Local government officials receive alerts about submitted issues and give prompt responses. SignalGuru [56] leverages smartphones to opportunistically detect current traffic signals with their cameras, collaboratively communicate and learn traffic signal schedule patterns, and predict their future schedule. Pedestrians distracted by smartphones are easy to meet with various dangers. Existing works about pedestrian safety are mostly based on the sensing capabilities from a single device. The surrounding information that can be learned, however, is quite limited or incomplete. CrowdWatch [57] leverages mobile crowd sensing to characterize fine-grained nearby contexts and prompt users in dangerous situations. Environmental protection is another topic that benefits from crowd photographing. For example, CreekWatch [5] allows volunteers to report information about waterways in order to aid water management programs. Jam Eyes [58] uses cameras of drivers who can observe the causes of a jam (e.g., a broken-down truck in the middle of the street) and shares the pictures or short videos with drivers in the jam line. WreckWatch [9] allows bystanders and uninjured victims to take pictures using their smartphones and share them with first responders after the car crash happened. SakuraSensor [59] automatically extracts flowering-cherry routes information from videos recorded by car-mounted smart-phones and shares the information among citizens. GigaSight [60] is a crowdsourced first-person video collection framework that can be employed for lost-object finding and public safety management.

V. Research Challenges and Key Techniques

In addition to the general issues of MCS systems such as incentives and task allocation, VCS has the following particular issues to be addressed. We present them in line with the working process of a VCS system. We also give a summary of the technical contributions of the major VCS applications described in Section IV, as shown in Table I.

A. Diversity-oriented Visual Task Allocation

Traditional MCS task allocation is based on point coverage [61, 62]. In contrast, VCS tasks are more about diverse aspect coverage and should consider multi-dimensional contexts in task allocation. For example, we should select workers from diverse directions and distances for a better characterization of an event in a VCS-based event sensing task. In other words, the optimization goal is to increase diversity in crowd-contributed data. A possible solution is to “decompose” a VCS task into a number of simple tasks (e.g., tasks with point coverage) according to the task constraints and human spatio-temporal distribution, and then allocate the tasks to the selected workers. Mobility prediction is important in task decomposition as we can use it to estimate the potential “point coverages”. Representative studies on human mobility prediction in mobile crowdsensing are investigated in [61, 63, 64].

Most VCS tasks are about static objects (e.g., SmartEye [41], SmartPhoto [25]). There are also dynamic sensing targets which have not been studied. Therefore, beyond “detection”-oriented VCS sensing tasks, “tracking” becomes another type of VCS tasks. For example, when a terrorist event occurs, observers may report a suspect vehicle to be tracked. We should use visual techniques to measure the context of the vehicle, such as moving direction, speed, and its multi-view appearance.

The diversity needs can be represented as various task constraints. However, sometimes the various aspects of a task are difficult to determine as the task requesters are not familiar with the target or the constraint set cannot align well with the sensing contexts. In such cases, we may ask the task requesters to simply specify how many pictures they want to select from. Originally contributed data can be grouped and outliers (or noise data) can be filtered out by using crowd intelligence, as discussed in our previous work [65].

Various human-companioned mobile devices can be employed for sensing, including wearables, smartphones, vehicles, et al. This results in the device heterogeneity issue. For example, different devices have different capability on image/video quality and diverse network connections (e.g., 2G/3G/4G, WiFi). Due to this highly dynamic nature, modeling and predicting the sensing capabilities of each node to accomplish a particular task is difficult. When there are a large number of available devices with diverse sensing capabilities, scheduling sensing and communication tasks among them under resource constraints will become more challenging [66].

B. Incentives and Participant Reliability

Incentive is a challenge to the human involvement in VCS. Without strong incentives, individuals may not be willing to participate in the sensing task with cost of their own limited resources. General purpose incentive mechanisms for MCS systems are reviewed in [67, 68]. There are additional requirements when designing incentive mechanisms for VCS systems. For example, it is crucial to guide people to capture pictures fulfilling the diversity needs of tasks. To motivate people to contribute data at specific places in MCS, [69, 70] displayed the rewarding points to users on the digital map. Their methods are related to point-coverage and cannot address the multi-dimensional coverage needs. As demonstrated by the studies such as PhotoCity [8], a well-designed user interface is important and can steer participants to attain high-quality sensing. For example, we can share with the participants detailed picture collection and payment dynamics, including the pictures collected by each participant, their shooting context and data quality, and payment results. Furthermore, Kawajiri et al. [70] use a point calculation method, where rewarding points for a place can be adjusted by learning crowd behaviors. This inspires us to develop adaptive utility measurement schemes, which may better steer people to cover less-popular aspects of a task.

Existing monetary-based incentive studies (e.g., the reverse auction based methods) mainly encourage user participation, whereas sensing quality is often neglected. The reliability of recruited workers (e.g., sensing capabilities, and uncontrollable mobility) should also be considered. There are several potential ways to address this. First, it is important to build worker models that can characterize a worker from different aspects, such as skills, experiences, interests, mobility, and reputation. The model can be applied in task allocation to estimate participant reliability and select appropriate workers. For instance, Zhang et al. [71] use worker confidence to estimate the reliability of successfully completing the assigned sensing tasks, and study the maximum reliability task assignment under a recruitment budget. Cheng et al. [33] estimate worker capability and assign workers to visual spatial tasks (e.g., taking videos/photos of a landmark or firework shows) such that the completion reliability and the spatial/temporal diversities of spatial tasks are maximized. Second, to ensure the reliability of crowdsourced data, we can recruit 'redundant' workers to perform the same task and further aggregates their sensing reports for truth discovery, as demonstrated by [72] and [73]. Third, it is also useful to integrate data quality measurement in the incentive mechanisms to motivate high-quality task completion. For instance, TaskMe [65] leverages a combination of multi-facet quality measurement and a multi-payment enhanced reverse auction scheme to improve sensing quality.

C. Data Selection and Redundancy Elimination

One critical issue in VCS is data redundancy. Redundant data should first be grouped, and then representative picture(s) from each group should be selected for further processing. In [74], a formal task model is defined, and the requirements on data redundancy are predefined as task constraints. For example, the view directions are either single (e.g., object price [49, 50]) or multiple (e.g., an event [34, 13]), and the status of the target might change slowly (e.g. posted fliers [10]) or quickly (e.g. traffic signals [56]). A brief summary of the task constraints and relevant applications is given in Table II. In view of this, both the data grouping and selection process of VCS should adapt to the various task requirements.

TABLE II
SELECTION CRITERIA AND RELATED VCS APPLICATIONS.

Task constraints	Representative Applications
Multiple shooting angle	SmartPhoto [25], PhotoCity [8]
Single shooting angle	FlierMeet [10], LiveCompare [49], PetrolWatch [51]
Change slowly	SmartPhoto [25], PhotoNet [7]
Change quickly	SignalGuru [56], WreckWatch [9]
Long/short shot distance	InstantSense [34], TaskMe [65], MoViMash [36]

There have been numerous studies on designing data selection schemes for VCS. For example, CrowdPic [24] proposes a generic picture collection framework that supports efficient picture grouping and redundancy elimination based on multi-dimensional task constraints. The pyramid-tree (PTree) algorithm is proposed to represent the task constraints and provide support for online crowdsourced picture grouping. Some VCS applications try to learn data selection strategies from human experience or professional knowledge. For instance, MoVieUp [37] incorporates a set of computational domain-specific filming principles summarized from a formal user study, e.g., the less shot switching principle and the 30 degree rule (there should be at least 30 degrees' difference between shooting angles) to avoid jump cuts in camera selection. MoViMash [36] is a framework that summarizes the video clips about an event from different shooting angles and distances. They have built a hidden Markov model to learn the experiences (e.g., decision making for shooting angle and distance selection, shot length and transitions) from professional editors.

Another issue regarding data selection is that we should filter out noisy or irrelevant data. A simple and straightforward hypothesis is that if more participants report an observation, it is more likely that the observation is relevant, whereas objects with few observers can be treated as outliers. However, isolated pictures are not always irrelevant, and the relevant sensing targets may locate in the places with few observers. There have been several studies that address this issue. In TaskMe [65], data utility or usefulness is measured by predefined task constraints while not by the clustering results. In PhotoNet [7], a picture is treated as an outlier, if it is geographically collocated with a popular picture cluster, but is visually different from the group. Otherwise, the singleton item is viewed as a rare item that is useful but has few observers.

Generally speaking, data selection is conducted offline at the server side. However, sometimes it should be done online in the client-server data transmission process, as discussed in the next subsection.

TABLE III
COMMUNICATION REQUIREMENTS IN VCS SYSTEMS.

Name	Requirement	Solution
CARE[40]	Low computation cost	Quality-based data selection
PhotoNet[7]	High data utility under limited storage capacity	Data selection based on spatio-temporal and visual difference
SmartPhoto[25]	High data utility	Data selection based on sensing context difference
SmartEye[41]	High data utility	Data selection based on sensing context difference
CooperSense[39]	Low computation cost	Opportunistic collaboratior and data selection
Piggyback crowdsourcing[76	Energy-efficient data transmission	Piggyback crowdsourcing
EMC^3[77]	Energy-efficient data transmission	Participant behavior prediction-based task assignment
Xiao et al. [78]	Energy-efficient data transmission	Static node cooperation
EnUp [79]	Energy-efficient data transmission	Networking condition prediction
Sun and Liu [80]	Load-balancing and reliable communication	Congestion-aware D2D-enabled incentive mechanism
Dong et al. [81]	Reliable and energy-efficient communication	Representative node selection mechanism
Wu et al. [82]	Reliable data transmissior	Hybrid routing scheme

D. *Opportunistic Visual Data Transmission*

Due to the limitations of communication bandwidth, storage and processing capability, it is a challenge to transfer the huge amount of crowdsourced pictures. Delay tolerant networks (DTNs) [75] have been proved a promising way to deliver data in poor network environments. However, even with DTN, how to save networking resources still poses numerous challenges. To attain efficient and timely delivery of crowdsourced pictures, the primary issue is to determine the value of the pictures based on their significance and redundancy, and only upload those valuable ones. As discussed earlier, a good visual coverage usually requires multiple views of the sensing target. We thus should measure the utility of a picture considering the unique aspect(s) it covers. The measured picture utility should be used as the inputs of data transmission protocols to enable efficient visual data transmission.

CARE [40] leverages image similarity detection algorithms to eliminate similar-looking pictures in picture delivery. Three state-of-the-art computer vision algorithms, including SIFT, pHash and GIST, are applied to balance the tradeoff between accuracy and computational cost. PhotoNet [7] is a picture delivery service that prioritizes the transmission of pictures by considering the spatio-temporal and visual difference. It aims to solve the diversity optimization problem by choosing a subset of objects whose total coverage is maximized, subject to some aggregate resource constraints (e.g., storage capacity). Wu et al. [20] propose a resource-aware photo crowdsourcing framework in DTN, which uses picture contexts such as location, orientation to build a photo coverage model and estimates picture utility. A photo selection algorithm is proposed to maximize the value of selected pictures, considering both *point* and *aspect* coverage. SmartPhoto [25] quantifies the quality of crowdsourced pictures based on the accessible geographical and geometrical info including the smartphone's orientation, position, and all related parameters of the built-in camera. Both the Max-Utility problem and Min-Selection problem are studied and greedy algorithms with theoretical performance bounds are proposed. SmartEye [41] implements QoS-aware in-network deduplication based on the software-defined networks (SDN). Two optimization schemes are developed, namely semantic hashing and space-efficient filters. CooperSense [39] proposes a local smartphone cooperation method to identify unique and high quality data for transmission. A summary of the communication requirements and relevant solutions of the major VCS systems discussed in this paper is given in Table III.

E. *Energy-Efficient and Reliable Communication*

Participation in VCS systems can easily expose users to a significant drain on limited battery resources of users' mobile devices. To maintain large-scale user participation, VCS system designers should minimize the energy consumption mainly due to the data transmission process between mobile clients and the backend server. Piggyback CrowdSensing (PCS) [76] is an energy-efficient MCS system that can intelligently leverage the opportunities for data collection that frequently occur during everyday smartphone user operations, such as placing calls or using applications. The EMC^3 framework [77] reduces energy consumption in data transmission by incorporating human behavior prediction (e.g., calls and human mobility) and intelligent task assignment. [78] propose a static-node-assisted data transmission protocol to attain energy-efficient opportunistic data transmission in crowdsensing systems. By forecasting network connections and smartphone usage, [79] intelligently schedule the data transmission process to minimize the overall energy consumption.

Reliability is another crucial requirement when deploying VCS systems in the real world. For example, the communication performance of crowdsensing may deteriorate in some high-density areas (e.g., shopping malls, and central business district streets) due to the overwhelming communication requests, whereas the wireless bandwidth in other areas may not be fully utilized with infrequent communication requests. To address this, [80] proposes a congestion-aware D2D (device to device)-enabled incentive framework to achieve efficient load balancing and provide real-time reliable communications in mobile crowdsensing. [81] studies the reliability and energy efficiency requirements as a whole and proposes a node-selection-based event data collection approach to meet both needs. [82] presents a hybrid routing scheme in vehicular networks for inter-vehicle, vehicle-to-roadside and inter-roadside data dissemination in urban hybrid networks, which can guarantee the reliability of data dissemination under various networking environments.

F. Lightweight Image Matching and Processing

As is for visual crowdsensing, image processing and computer vision techniques are indispensable to VCS. However, to increase the efficiency on processing large-scale crowdsourced pictures, lightweight and robust computer vision techniques should be introduced. We group diverse VCS tasks into image matching, 3D modeling, image tagging, and text/sign recognition, as discussed below.

Image matching is frequently used in VCS for redundancy detection [7, 24] or reference object identification (e.g., store logos, information desks) in visual-based localization or navigation [44, 46]. In image processing, there are two popular image feature vector extraction algorithms, namely SIFT (Scale Invariant Feature Transform) and SURF (Speeded Up Robust Features) [22]. Experiments show that SURF is much faster while achieving comparable accuracy to SIFT [30]. According to this finding, CrowdMap [27] uses Histogram of Oriented Gradients (HOG) [83] descriptor computing algorithm to select key video frames and then uses SURF for efficient image matching. Other methods for fast image matching are also studied. For example, Travi-navi [48] adopts the ORB algorithm as it is faster than SURF and SIFT and can extract image features in real time on mobile devices. CrowdPan360 [31] represents an image with a short bit string (called image fingerprinting), which can capture the perceptual features of the image. They use the perceptual hash algorithm [84] to generate fingerprints. To accelerate image matching, picture grouping is often used. FOCUS [35] compares the geometric (line-of-sight) relationship between the content of videos. A strong geometric overlap in a pair of videos indicates that they both capture the same subject.

3D modeling that reconstructs scenes or views of an environment has been widely used for indoor mapping and localization. Most of current SLAM (simultaneous localization and mapping [85])-based indoor scene reconstruction techniques (e.g., Google Cartographer[2] and Xsens Scannect [86]) require specialized equipment to capture indoor scenes and have poor scalability. Different from SLAM, Structure-from-Motion (SfM) techniques [87] enable 3D modeling of surrounding environments using unordered 2D pictures. A typical SfM pipeline includes three steps: feature extraction, feature matching, and bundle adjustment. Highly distinctive features are first extracted from images using algorithms like SIFT. Image matching is then conducted over the features between image pairs. The matches are finally used as the input for the bundle adjustment component for producing optimal estimates of camera poses and the locations of 3D points. The typical implementations of SfM include VisualSFM [88] and Bundler [89]. Based on SfM, Agarwal et al. [90] construct 3D models of Rome from 150K photos found from Internet photo sharing sites. [47] [35] [28] take crowdsourced photos as the input to build 3D models of the indoor space of interest using SfM techniques. To decrease computation load in SfM, iMoon [47] introduces density-based 3D model partitioning and fingerprint-based partition selection.

Image tagging facilitates grouping and browsing of crowdsourced pictures. DietSense [53] explores standard image processing techniques, including dominant color analysis [91] and histogram Kullback-Leibler (K-L) divergence [92], to tag and cluster crowdsourced food-pictures. For example, the dominant color analysis of images is effective for place tagging, e.g., pictures that are primarily blue or green largely correspond to outdoor environments.

Photos taken from real-world environments usually contain signs and descriptive texts. Such information is useful for a number of VCS tasks, such as place categorization and localization, by applying sign recognition and optical character recognition (OCR) techniques to extract information from pictures. For example, CrowdSense@Place [46] uses a commercial OCR engine to extract written texts from posters or signs within places. In [31], Microsoft's stroke width transform algorithm [93] is used for identifying texts (e.g., departments, cafeteria, and street names) in crowdsourced images.

G. Picture Quality Estimation

Although VCS provides a cheap way of collecting pictures of interesting targets, there are always uncertainty issues regarding the quality of user-contributed data. For example, a user-captured picture can be blurry or has undesired brightness. The target may also be blocked by unexpected obstacles. Therefore, we should estimate the quality of crowdsourced pictures and eliminate low-quality ones. There are several potential ways to estimate data quality.

Content-based quality estimation. MoViMash [36] develops an edge-density based method to detect videos with occluded views. It is based on the assumption that the pictures with object occlusions will result in lower edge density than the original one in event sensing. Similarly, Travi-Navi [48] uses the number of detectable ORB features in images as the quality metric to quantify the image quality in terms of blurs. DietSense [53] employs the Roberts cross edge detection algorithm [94], where "edgy" pictures (by counting the number of computed black pixels) were filtered as they may contain homogeneous environments (e.g., walls, empty desks, pictures of the floor). SmartPhoto [25] uses Depth-of-Field (DOF) to determine if the target is out of focus. If the target falls into the DOF, the photo is considered valid.

Context-based quality estimation. Movi [13] selects videos with a good view that have high accelerometer reading rankings and light intensity detected by embedded sensors in smartphones. FlierMeet [10] proposes an approach that uses crowd intelligence to determine the best shooting angle to the target (e.g., fliers posted on boards). To deal with blurry images

[2] https://github.com/googlecartographer

caused by vehicle vibrations in [51], a set of pre-selection thresholds based on the measures from embedded accelerometer of the mobile phone are designed.

Hybrid feature-based quality measurement. There are also studies that try to integrate content and context features to augment quality estimation in VCS tasks. In IndoorCrowd2D [30] and Travi-Navi [48], real-time data quality feedback mechanisms are implemented to guide users to provide high quality data. The metrics are measured by processing the sensor data and the image data in real time, including linear acceleration, angular acceleration and the number of SURF features in each picture. If the prior two values are beyond a certain threshold, it indicates that the user either moves or turns too fast. If the number of SURF features falls below a predefined threshold, this exhibits that the user shoots feature-less objects, such as walls.

H. Large-Scale Visual Data Understanding

The inherent nature of crowdsensing makes it challenging to analyze and understand large-scale crowdsourced data. To ensure efficient visual data mining and understanding, there are two potential research directions.

Novel machine learning methods. Large-scale image classification has recently received significant interest from the computer vision and machine learning communities. Several large-scale visual data sets have been created. For instance, ImageNet[3] consists of more than 14M images labeled with almost 22K concepts [95], and the Tiny image data set consists of 80M images corresponding to 75K concepts [96]. In their pioneering work, Lin et al. [97] employ high-dimensional image descriptors in combination with linear classifiers to ensure computational efficiency in large-scale image classification. In the survey paper about the ImageNet data challenge Large-Scale Visual Recognition Challenge (ILSVRC), Russakovsky et al. [98] review the novel methods developed regarding the large-scale data mining tasks. Akata et al. [99] benchmark several SVM objective functions (e.g., one-versus-rest, ranking, and weighted approximate ranking) for large-scale image classification over the ImageNet data set. They find that one-versus-rest is simple and can be easily parallelized to address the large-scale data processing issue, and by using SGD (stochastic gradient descent)-based learning algorithms, ranking-based approaches can also scale well to large data sets. Deep learning methods, such as Convolutional Neural Networks (CNNs) have also been demonstrated as an effective class of models for large-scale image content understanding [100]. It has also been demonstrated useful when applied on large-scale video classification, using a new data set of 1 million YouTube videos belonging to 487 classes [101].

The integration with crowd intelligence. For many problems about image understanding, humans can still perform more accurately and efficiently than a machine. We notice that the knowledge hidden in the process of data generation, regarding individual or crowd behavior patterns are neglected in crowdsourced data mining. We intend to address the challenge from a new perspective: harnessing the power of crowd intelligence to better understand large-scale crowdsourced data. There are several representative studies that use crowd intelligence. FOCUS [35] leverages shared content recognition by the overlap of line of sight to guide view selection in crowdsourced video mashup. Movi [13] reports how to use crowd behavior patterns to identify potential social interests in a social activity (e.g., a party). It designs two types of human intelligence, including specific event signatures (e.g., laughter, clapping, and shouting) and group behavior patterns (e.g., group rotation, acoustic-ambience fluctuation). Note that the usage of crowd intelligence does not necessarily replace the role of image processing algorithms, but is to serve as an important complement to improve the utility of the collected photos, especially when resources are constrained.

VI. Insights and Future Directions

Having presented the challenges and key techniques developed to addressing various issues in VCS, this section discusses our insights for the future research directions of VCS.

A. A Generic VCS Framework

Existing VCS systems usually only support one specific task (e.g. river pollution monitoring [5] and disaster/event picture collection [7, 40, 41]). This leads to reusability and scalability limitations as these systems are application-dependent. Regarding the challenges and techniques presented in the previous section, we have proposed a generic framework for VCS. The motivation for building a generic framework for VCS is inspired by MTurk [17], and has the following merits.

First, this framework facilitates the rapid specification of VCS tasks taking into consideration different constraints, eliminating the need to develop domain-specific, application-dependent proprietary systems. Second, it lowers the barrier for regular users to publish VCS tasks and meet their personalized needs. Third, it provides mobile users with a unique way to access VCS tasks, which can simplify participant recruitment and data consumption.

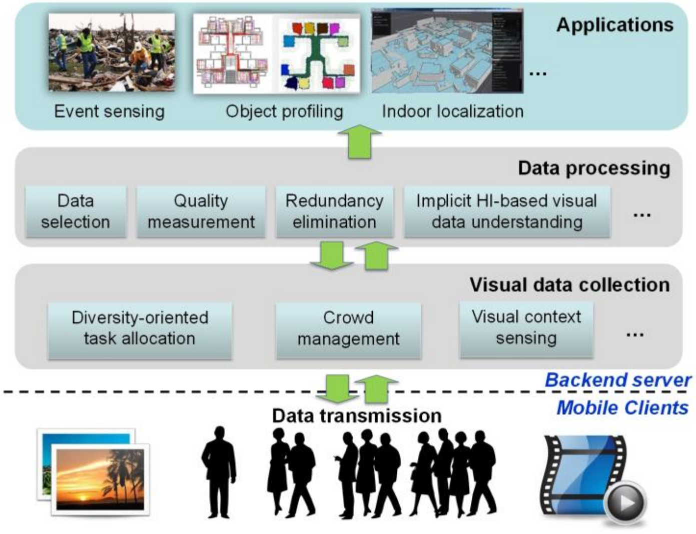


Fig. 5. A generic VCS framework

[3] http://www.image-net.org

The layered client-server architecture of the framework is illustrated in Fig. 5. At the *mobile client* side, visual content and video clips are captured by the crowd according to task needs. The *data transmission* layer defines low-cost client-server transmission protocols that are particularly important for visual data collection using smartphones. The backend-server side incorporates three layers: the visual data collection layer, the data processing layer, and the application layer.

- ***Visual data collection***. *Diversity-oriented task allocation* is responsible for task decomposition and assignment regarding various task constraints and spatio-temporal participant distribution. *Crowd management* maintains the profile and real-time contexts of participants. *Visual context sensing* extracts contexts about crowd-object interaction (e.g., picture shooting).
- ***Data processing***. The *data selection* component is responsible for selecting representative data according to task constraints. The *redundancy elimination* module groups crowdsourced pictures and identifies redundancy at both content and semantic levels. The *quality estimation* component filters low-quality pictures with context- and content-based techniques. Finally, to facilitate large-scale data understanding, various types of crowd intelligence are extracted from crowd-object interaction patterns.
- ***Applications***. It makes use of the high-quality crowd-contributed data in various application areas, as the ones presented in Section IV.

It should be noted that VCS are essentially crowd-powered *mobile* camera networks. Though there are quite a few differences, we can still learn much from well-studied *static* camera networks [102]. For example, many challenges discussed in this paper, including low-cost communication [103], sensing coverage optimization [104], and efficient visual data processing [105], have also been investigated in traditional camera networks. Furthermore, the sensing capabilities of static and mobile camera networks are often complementary. For example, regarding event sensing, stationary cameras can deliver high-quality, near real-time data, while mobile cameras can significantly enhance event sensing coverage and data diversity. In other words, different types of sources can capture different aspects of a sensing target, and thus complementary data should be collected from each source to generate a complete picture.

As a promising extension of the VCS framework, we should investigate the integration of stationary camera networks with VCS, i.e., the building of collaborative sensing systems with both pre-deployed cameras and mobile cameras. In D-CPSS [106], a collaborative sensing layer is incorporated in the proposed data-centric framework for cyber-physical-social systems. It is used to manage the scheduling and cooperation of the selected sensing sources according to the dynamics of the sensing task. [107] studies the full view coverage problem in heterogeneous camera networks, i.e., a combination of stationary and mobile camera networks. The collaboration of mobile and static sensing nodes can also contribute to high-performance data transmission. For example, [108] propose a static-node-assisted adaptive data dissemination in vehicular networks, which can be used to lower data dissemination latency. [78] also investigates the deployment of static nodes to enable energy-efficient data transmission in crowdsensing.

B. Embracing Mobile Edge Computing

When deploying large-scale VCS systems in the real-world, we should particularly consider communication issues, such as scalability and delivery latency. For example, for crowdsourced video collection, a key challenge is the high cumulative data rate of incoming videos from many users to the backend server (in the cloud). Without careful system design, it could easily overwhelm the capacity of networking paths to the centralized cloud infrastructure, considering that 12,000 users transmitting 1080p video would require a link of 100GB per second.

Mobile Edge Computing (MEC) is a new paradigm that reforms the cloud hierarchy by placing computing resources, referred to as *cloudlets*, at the Internet's edge in close proximity to mobile devices [109,110]. MEC has been recognized by the European 5G PPP (5G Infrastructure Public Private Partnership) research body as one of the key emerging technologies for 5G networks [111]. The major aims of MEC are to reduce latency, ensure scalable network operation and service delivery, and offer an improved user experience. There are several merits by integrating MEC with VCS systems, some of which are closely related to the aforementioned research challenges.

First, the cumulative network-bandwidth demand into the cloud from a large collection of high-bandwidth mobile cameras can be considerably lowered, if the raw data is analyzed on cloudlets and only the extracted information or metadata [112,113] is transmitted to the cloud. Simoens et al. [60] propose a scalable system for continuous collection of crowdsourced videos from mobile devices. It achieves scalability by decentralizing the collection infrastructure using cloudlets based on virtual machines.

Second, the privacy issue can be mitigated. By serving as the first point of contact in the infrastructure, a cloudlet can enforce the privacy policies of its owner prior to the release of the data to the cloud [109]. A user should be able to delete or denature a subset of sensor data she deems sensitive [60,114]. Denatured sensor data becomes safe to release, e.g., faces in images can be blurred, sensor readings can be coarsely aggregated, etc.

Third, real-time context-aware computing (e.g., human behavior/mobility prediction, the sensing context learning) is important in VCS systems. This is challenging when running on resource-constrained mobile devices. MEC can facilitate efficient context-aware computing by allowing mobile devices to outsource their computation to the upper-layer cloudlets [115].

C. Augmented Data Understanding with Crowd Intelligence

We have presented the usage of crowd intelligence to facilitate large-scale crowdsourced data understanding. It is useful for at least the following task types, and a summary is given in Table IV.

- *Data filtering*. Filtering our noisy or low quality data.

- *Data classification and tagging*. Categorizing the data or assigning tags to the data.
- *Data clustering & segmentation*. Grouping redundant data. For evolutionary objects such as events, it is often important to segment the data stream.
- *Data selection*. Selecting representative data from the redundant data set.

TABLE IV
UNDERSTANDING OF CROWDSOURCED DATA WITH CROWD INTELLIGENCE.

Task type	**Related work**	**The usage of Crowd Intelligence**
Filtering	Data quality measurement [10]	Aggregated shooting behaviors
Classification & Tagging	Flier tagging [10], Place categorization [46, 116]	Group structure, Crowd-object interaction patterns
Clustering & Segmentation	Highlight detection [13], Subevent detection [34]	Group behavior patterns (rotation, laughing), Individual/Crowd photographing patterns
Data selection	Redundancy elimination [24], Event summary [34]	Picture shooting contexts

Crowd intelligence can be applied directly or indirectly for understanding crowdsourced data. When used *directly*, it often acts as the parameter input of a decision making function (e.g., data selection or filtering). For example, FlierMeet [10] use the central-tendency of crowd picture shooting angles as the parameter input of a data-filtering function for picture quality measurement. Movi [13] use group behavior patterns to identify potentially interesting scenes in social events. When used *indirectly*, crowd intelligence is normally integrated with MI, via data mining or machine learning algorithms. It can be used as important features of machine learning algorithms (e.g., clustering or classification methods). For example, crowd shooting patterns have been used for event segmentation in InstantSense [34]. Crowd-contributed visual cues can be used to recognize the ambient contexts of places [116].

There are several interesting directions to be investigated further in the future, as discussed below.

- ***The scope of crowd intelligence.*** The major types of crowd intelligence presented in this paper include crowd behavior patterns, crowd-object interaction patterns, and so on. Crowd intelligence has a wide scope in terms of cognitive abilities, individual attributes, social features, interaction and behavior patterns. It is crucial to characterize them and investigate their usage in crowdsourced data mining.
- ***The emergence of crowd-machine computational systems.*** Crowd intelligence is used as feature inputs for machine learning and data mining algorithms. With the manifold efforts of embedding human intelligence in computing systems, we will finally build crowd-machine computational systems. The complementary features of crowd and machine intelligence should be further explored and new integration or collaboration manners should be studied.

D. Unique Privacy Issues

The VCS data consists of the participant's context and the visual content. The former one mainly exposes the participant's privacy, which is similar to other forms of MCS apps. However, compared to the other types of crowdsensed data, visual contents in VCS can expose both participant's and the passerby's privacy. The privacy information exposed by leveraging context learning and visual content understanding may include human's location, identification, occupation, activity, hobby, etc. We first characterize the two diverse privacy concerns below.

(1) Participant privacy. Privacy leakage concern is one of the problems that prevent people from participating in VCS tasks, which we call the participant privacy concern. To this end, most VCS tasks use monetary rewards in return for people to contribute data. To motivate user participation, we should also explore new techniques to protect personal privacy while allowing their devices to reliably contribute data. One such effort is the AnonySense architecture proposed by Cornelius et al. [117], which supports the development of privacy-aware applications based on crowd sensing. Other techniques on participant privacy protection in crowdsensing have also been reviewed in [118].

(2) Third-person privacy. People and objects in public areas can be unintentionally captured by VCS task workers, which can reveal the privacy information beyond the picture-taker. In some emergency (e.g., disaster relief or public safety) or social (e.g., a party) occasions, people might not be that sensitive to data privacy because we have trusted picture-takers and 'controllable' or 'predictive' data usage. For instance, Movi [13] assumes that attendants in a social party may share mutual trust, and hence, may agree to collaborative sensing and data sharing. However, in other occasions, we should protect the privacy of the passersby and other sensitive objects. There are at least two critical issues to be addressed, regarding how to identify the sensitive information and how to avoid the exposure of it. Though there are still not technical standards for dealing with these issues, there are several promising methods to be leveraged, as discussed below.

- ***Intentional image blurring.*** People's face and vehicles' plate number are usually sensitive information in images. One common method used for privacy protection in visual systems is to blur certain parts of images. Google street view[4] blurs the faces detected in the collected visual items for outdoor scene reconstruction. GigaSight [60] uses denaturing to protect the privacy of people in videos, such as blurring all faces or only a subset of faces from a given list. The referred computer vision techniques to enable this include face detection, face recognition, plate number recognition, and object recognition in individual frames or images. Beyond faces and vehicle plate number that have common consensus from people, there can be other sensitive information, such as special human activities, brands, and sensitive sites. It is difficult to pre-define such situation-specific privacy objects. Domain knowledge or

[4] http://www.google.cn/maps/

human efforts are often needed to address these issues.

- ***Non-visual information extraction***. Some applications need the visual information of sensing targets (e.g., flower blooming [59], events [34, 13]), while sometimes we only need the semantic information extracted from the pictures (e.g., object prices [50], traffic signals [56]). Therefore, it is not always necessary to deliver complete pictures to task requesters. For example, MobiShop [50] extracts texts on shopping bills using the OCR technology. In such cases, image processing can be conducted at the client or server side and only the information distilled should be delivered to the task requesters. This can rely on commonly-used image processing techniques, such as object recognition, OCR, image tagging.

Beyond these discussions, it is also important to refer to existing solutions in visual sensor networks (e.g., camera networks) when addressing vision-related data privacy issues. A thorough survey has been given by Winkler et al. in [119].

E. Field Study and Experiments

As a crowd-driven research field, how to conduct experiments to validate the techniques/approaches is a challenge. We first make a summary of the existing methods used for VCS evaluation, as shown in Table V.

From the summary given in Table V, we can derive the following conclusions and guidelines for evaluation of VCS-related techniques.

- ***Combined manners for evaluation***. Crowdsensing by recruiting large-scale participants is of high cost. Therefore, we can find that most VCS studies have only limited participants. Therefore, simulations are usually employed for large-scale studies. Though, there are many parameters in visual crowdsensing that are difficult to simulate, and there still lacks a generic simulation tool for VCS research. Compared to simulations, field studies with real-world deployments can better validate the effectiveness of the methods/techniques used and identify the problems that are not easy to be found in experimental environments. To demonstrate the robustness and usability of the methods in different environments, some works conduct two more field studies [30, 13, 36, 44].
- ***Long-term, large-scale field studies***. Though a few VCS studies have chosen to conduct field studies in buildings [30, 47], shopping malls [44], or university campus [10] to validate their system, the scale of these studies is still limited due to the high cost. In the future, it is better to publish the tasks as smartphone applications to have wide participation of people. For online crowdsourcing, there have been studies for testing in commercial platforms, such as MTurk [17]. There have also been some startups of MCS platforms (e.g., Ohmage[5]), and it is promising to conduct experiments by collaborating with such platforms.
- ***Leveraging online crowdsourced data***. Though it is difficult to collect real-world data sets with a large number of participants, we find that several studies [34, 37] have employed the visual resources (pictures, videos) from the social websites (e.g., Flickr, YouTube). The benefit of using online resources is obvious as *i*) we can easily obtain rich open data [121-123], and *ii*) they are also crowdsourced resources and maintain major features of visual crowdsensing. However, online resources are mainly in contents while the metadata such as shooting contexts (e.g., shooting angle, shakiness) are not attached. Therefore, for some tasks that have multi-dimensional constraints, it should synthesize with simulation-based methods [7, 25].

TABLE V
EXPERIMENT SETUP IN EXISTING VCS STUDIES.

Name	**Method**	**Participants and Data set**
Wu et al. [20]	By simulation, randomly generate photos over real traces	Mobility traces from Mixed Realit (100 people) and Cambridge06 dat sets (36 people)
SmartPhoto [25]	A prototype App, Real world demo (100m*100m area) + extensive simulations	30 photos of a building for the demo; Tens of targets randomly distributed, and thousands of "virtual" photos for simulation
Xu et al. [44]	Field study in a shopping mall and a food plaza	More than 1000 photos are taken for the 50 and 41 POIs in the shopping mall and the food plaza
iMoon [47]	Field study in a real building (1100 m^2)	3D modeling generated from 2,197 pictures
IndoorCrowd2D [30]	Field study in two buildings (a teaching building and a GYM)	25 participants, 55,453 pictures contributed
PhotoNet [7]	Simulation over ONE[120] for a post disaster rescue mission in town	100 virtual participants, 1000 real pictures of different landmarks in a campus randomly tagged to the participants
CARE [40]	Simulation over ONE[120] over a town area for a disaster scenari	50 virtual participants randomly located inside the simulated disaste area
CrowdPAN360 [31]	Field study in a campus, five-week period	10 participants, 6,000 more crowdsourced pictures about 70 indoor/outdoor objects
FOCUS [35]	Field study in a football stadium, a two-month period	70 participants, 325 video streams and 412 pictures
Movi [13]	Field study over a thanksgiving party and a smart-home tour	A total of 21 participants
InstantSense [34]	Field study over two events in a campus and synthesized study with 7 online videos	A total of 328 event pictures by 21 participants; Seven online event videos from Youku and eight participants for tagging
FlierMeet [10]	Field study in a campus, a eight-week period	38 participants, 2,035 pictures about 921 objects in the campus
MoViMash [36]	Field study for three publi performance events	A total of 29 participants for video recording and 17 participants for user study
MoVieUp [37]	Synthesized study with online resources	46 mobile recordings of six events collected from Youtube

[5] http://ohmage.org/, accessed: Feb. 4, 2016

VII. Conclusion

This paper has presented visual crowdsensing (VCS), an emerging research area that leverages regular users to photograph the interesting targets using their smart phones in the real world. We clarify the main characters of VCS, including the generic work flow of VCS tasks, the definitions of task coverage and data redundancy, crowd-object interaction contexts in data collection, as well as the triple concept graph. We have made a summary and comparison of different types of VCS applications, including object profiling, dynamic event sensing, indoor localization or navigation, disaster relief, personal wellness, and urban sensing. The unique challenges faced by VCS as well as the main techniques/solutions are further studied, such as diversity-oriented task allocation, data selection and redundancy elimination, opportunistic visual data transmission, energy-efficient and reliable communication, quality estimation, and visual data understanding. Based on the reviewing of existing systems and the identified characters, we have proposed a generic framework for developing VCS systems. We finally discuss our insights for the future research directions and opportunities of VCS.

There are several crucial and promising research directions of VCS. First, visual sensing can provide rich information regarding our working or living environments. Though there have been numerous attempts of leveraging the power of crowd to facilitate large-scale visual sensing, we believe that there are still various VCS-enabled applications that can be enriched, by integrating with different domains. Inspirations can be partly drawn from the existing studies in the image processing and computer vision community. Second, compared to other modalities of crowdsensing tasks, VCS faces many unique challenges and the study of some of them are still at the early stage. At least the following topics need further investigation: diversity-oriented visual task allocation, efficient visual data selection and processing methods, the embracing of Mobile Edge Computing techniques, the usage of crowd intelligence for visual data understanding, and third-person privacy protection schemes. Third, we anticipate the development and deployment of large-scale VCS systems in the coming years, which will help identify the practical issues and evaluate the performance of the proposed methods.

References

[1] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, “Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm,” *ACM Computing Surveys*, vol. 48, no. 1, p. 7, 2015.

[2] H. Ma, D. Zhao, and P. Yuan, “Opportunities in mobile crowd sensing,” *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, 2014.

[3] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M’hamed, “Sparse mobile crowdsensing: challenges and opportunities,” *IEEE Communications Magazine*, vol. 54, no. 7, pp. 161–167, 2016.

[4] C. Chen, D. Zhang, X. Ma, B. Guo, L. Wang, Y. Wang, and E. Sha, “CrowdDeliver: Planning city-wide package delivery paths leveraging the crowd of taxis,” *IEEE Trans. on Intelligent Transportation Systems*, 2016.

[5] S. Kim, C. Robson, T. Zimmerman, J. Pierce, and E. M. Haber, “Creek watch: pairing usefulness and usability for successful citizen science,” in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2011, pp. 2125–2134.

[6] S. Reddy, D. Estrin, M. Hansen, and M. Srivastava, “Examining micropayments for participatory sensing data collections,” in *Proc. of the 12th ACM international conference on Ubiquitous computing (Ubicomp)*. ACM, 2010, pp. 33–36.

[7] M. Y. S. Uddin, H. Wang, F. Saremi, G.-J. Qi, T. Abdelzaher, and T. Huang, “Photonet: a similarity-aware picture delivery service for situation awareness,” in *Proc. of the 32nd Real-Time Systems Symposium (RTSS)*. IEEE, 2011, pp. 317–326.

[8] K.Tuite, N. Snavely, D.-y.Hsiao,N. Tabing,andZ. Popovic, “Photocity: training experts at large-scale image acquisition through a competitive game,” in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2011, pp. 1383–1392.

[9] J. White, C. Thompson, H. Turner, B. Dougherty, and D. C. Schmidt, “Wreckwatch: Automatic traffic accident detection and notification with smartphones,” *Mobile Networks and Applications*, vol. 16, no. 3, pp. 285–303, 2011.

[10] B. Guo, H. Chen, Z. Yu, X. Xie, S. Huangfu, and D. Zhang, “Fliermeet: a mobile crowdsensing system for cross-space public information reposting, tagging, and sharing,” *IEEE Trans. on Mobile Computing*, vol. 14, no. 10, pp. 2020–2033, 2015.

[11] Y. Jiang, X. Xu, P. Terlecky, T. Abdelzaher, A. Bar-Noy, and R. Govindan, “Mediascope: selective on-demand media retrieval from mobile devices,” in *Proc. of the 12th International Conference on Information Processing in Sensor Networks (IPSN)*. ACM, 2013, pp. 289–300.

[12] X. Sheng, J. Tang, X. Xiao, and G. Xue, “Sensing as a service: Challenges, solutions and future directions,” *IEEE Sensors Journal*, vol. 13, no.10, pp. 3733-3741, 2013.

[13] X. Bao and R. Roy Choudhury, “Movi: mobile phone based video highlights via collaborative sensing,” in *Proc. of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 2010, pp. 357–370.

[14] S. Maharjan, Y. Zhang, and S. Gjessing, “Optimal Incentive Design for Cloud-Enabled Multimedia Crowdsourcing,” *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp.2470-2481, 2016.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo, Y., “Earthquake shakes Twitter users: real-time event detection by social sensors,” in *Proc. of the 19th International Conference on World Wide Web (WWW)*, 2010, pp. 851-860).

[16] N. Maisonneuve, M. Stevens, M.E. Niessen, and L. Steels, “NoiseTube: Measuring and mapping noise pollution with mobile phones,” *Information Technologies in Environmental Engineering*, pp.215-228, 2009.

[17] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2008, pp. 453–456.

[18] M.R. Ra, B. Liu, T. F. La Porta, and R. Govindan, “Medusa: A programming framework for crowd-sensing applications,” in *Proc. of the 10th International Conference on Mobile Systems, Applications, and Services (Mobisys)*. ACM, 2012, pp. 337–350.

[19] Z. Yu, H. Xu, Z. Yang, and B. Guo, “Personalized travel package with multi-point-of-interest recommendation based on crowdsourced userfootprints,” *IEEE Trans. on Human-Machine Systems*, vol. 46, no.1, pp. 151–158, 2016.

[20] Y. Wu, Y. Wang, W. Hu, X. Zhang, and G. Cao, “Resource-aware photo crowdsourcing through disruption tolerant networks,” in *Proc. of the 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2016, pp. 374–383.

[21] Y. Wang and G. Cao, “Barrier coverage in camera sensor networks,” in *Proc. of the 12th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*. ACM, 2011, p. 12.

[22] L. Juan and O. Gwun, “A comparison of sift, pca-sift and surf,” *International Journal of Image Processing (IJIP)*, vol. 3, no. 4, pp. 143–152, 2009.

[23] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, “Efficient color histogram indexing for quadratic form distance functions,” *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.

[24] H.Chen, B. Guo, Z. Yu, L.g Chen, X. Ma. "A Generic Framework for Constraint-Driven Data Selection in Mobile Crowd Photographing", *IEEE Internet of Things*, vol. 4, no. 1, 2017, pp. 284-296.

[25] Y. Wang, W. Hu, Y. Wu, and G. Cao, “Smartphoto: a resource-aware crowdsourcing approach for image sensing with smartphones,” in *Proc. of the 15th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc)*. ACM, 2014, pp. 113–122.

[26] D. Mizell, “Using gravity to estimate accelerometer orientation,” in *Proc. of 7th IEEE Int. Symposium on Wearable Computers (ISWC)*. Citeseer, 2003, p. 252.

[27] S. Chen, M. Li, K. Ren, and C. Qiao, “Crowd map: Accurate reconstruction of indoor floor plans from crowdsourced sensor-rich videos,” in *Proc. of the 35th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2015, pp. 1–10.

[28] R. Gao, M. Zhao, T. Ye, F. Ye, Y. Wang, K. Bian, T. Wang, and X. Li, “Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing,” in *Proc. of the 20th annual international conference on Mobile computing and networking (MobiCom)*. ACM, 2014, pp. 249–260.

[29] A. Sankar and S. Seitz, “Capturing indoor scenes with smartphones,” in *Proc. of the 25th annual ACM symposium on User interface software and technology (UIST)*. ACM, 2012, pp. 403–412.

[30] S. Chen, M. Li, K. Ren, X. Fu, and C. Qiao, “Rise of the indoor crowd: Reconstruction of building interior view via mobile crowdsourcing,” in *Proc. of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM, 2015, pp. 59–71.

[31] V. Raychoudhury, S. Shrivastav, S. S. Sandha, and J. Cao, “Crowd-pan360: Crowdsourcing based context-aware panoramic map generation for smartphone users,” *IEEE Trans. on Parallel and Distributed Systems*, vol. 26, no. 8, pp. 2208–2219, 2015.

[32] S. H. Kim, Y. Lu, J. Shi, A. Alfarrarjeh, C. Shahabi, G. Wang, and R. Zimmermann, “Key frame selection algorithms for automatic generation of panoramic images from crowdsourced geo-tagged videos,” in *Proc. of the International Symposium on Web and Wireless Geographical Information Systems (W2GIS)*. Springer, 2014, pp. 67–84.

[33] P. Cheng, X. Lian, Z. Chen, R. Fu, L. Chen, J. Han, and J. Zhao, “Reliable diversity-based spatial crowdsourcing by moving workers,” *Proceedings of the VLDB Endowment*, vol. 8, no. 10, pp.1022-1033, 2015.

[34] H. Chen, B. Guo, Z. Yu, and Q. Han, “Toward real-time and cooperative mobile visual sensing and sharing,” in *Proc. of the 35th Annual IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 2016, pp. 1–9.

[35] P. Jain, J. Manweiler, A. Acharya, and K. Beaty, “Focus: clustering crowdsourced videos by line-of-sight,” in *Proc. of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM, 2013, p. 8.

[36] M.K.Saini, R.Gadde, S.Yan, and W.T.Ooi, “Movimash: online mobile video mashup,” in *Proc. of the 20th ACM International Conference on Multimedia*. ACM, 2012, pp. 139–148.

[37] Y. Wu, T. Mei, Y.-Q. Xu, N. Yu, and S. Li, “Movieup: Automatic mobile video mashup,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 1941–1954, 2015.

[38] N. Frey and M. Antone, “Grouping Crowd-Sourced Mobile Videos for Cross-Camera Tracking,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 800-807.

[39] H. Chen, B. Guo, and Z. Yu, “Coopersense: A cooperative and selective picture forwarding framework based on tree fusion,” *International Journal of Distributed Sensor Networks*, vol. 2016, 2016.

[40] U. Weinsberg, A. Balachandran, N. Taft, G. Iannaccone, V. Sekar, and S. Seshan, “Care: Content aware redundancy elimination for disaster communications on damaged networks,” arXiv preprint arXiv:1206.1815, 2012.

[41] K. Atukorala, D. Wijekoon, M. Tharugasini, I. Perera, and C. Silva, “Smarteye integrated solution to home automation, security and monitoring through mobile phones,” in *Proc. of the 3rd International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST)*. IEEE, 2009, pp. 64–69.

[42] Z. Yang, Z. Zhou, and Y. Liu, “From RSSI to SCI: Indoor localization via channel response,” *ACM Computing Surveys*, vol. 46, no. 2, p. 25, 2013.

[43] L. M. Ni, D. Zhang, and M. R. Souryal, “RFID-based localization and tracking technologies,” *IEEE Wireless Communications*, vol. 18, no. 2, pp. 45–51, 2011.

[44] H. Xu, Z. Yang, Z. Zhou, L. Shangguan, K. Yi, and Y. Liu, “Enhancing wifi-based localization with visual clues,” in *Proc. of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2015, pp. 963–974.

[45] R. Gao, Y. Tian, F. Ye, G. Luo, K. Bian, Y. Wang, T. Wang, and X. Li, “Sextant:Towardsubiquitousindoorlocalizationservicebyphoto-taking of the environment,” *IEEE Trans. on Mobile Computing*, vol. 15, no. 2, pp. 460–474, 2016.

[46] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, “Automatically characterizing places with opportunistic crowdsensing using smartphones,” in *Proc. of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*. ACM, 2012, pp. 481–490.

[47] J. Dong, Y. Xiao, M. Noreikis, Z. Ou, and A. Yl ̈a-J¨ a¨aski, “imoon: Using smartphones for image-based indoor navigation,” in *Proc. of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM, 2015, pp. 85–97.

[48] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, “Travinavi: Self-deployable indoor navigation system,” in *Proc. of the 20th annual international conference on Mobile computing and networking (MobiCom)*. ACM, 2014, pp. 471–482.

[49] L. Deng and L. P. Cox, “Livecompare: grocery bargain hunting through participatory sensing,” in *Proc. of the 10th workshop on Mobile Computing Systems and Applications (HotMobile)*. ACM, 2009, p. 4.

[50] S. Sehgal, S. S. Kanhere, and C. T. Chou, “Mobishop: Using mobile phones for sharing consumer pricing information,” in *Demo Session of the Intl. Conference on Distributed Computing in Sensor Systems (DCOSS)*, vol. 13, 2008.

[51] Y. F. Dong, L. Blazeski, D. Sullivan, S. S. Kanhere, C. T. Chou, and N. Bulusu, “Petrolwatch: Using mobile phones for sharing petrol prices,” 2009.

[52] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White et al., “Vizwiz: nearly real-time answers to visual questions,” in *Proc. of the 23nd annual ACM symposium on User interface software and technology (UIST)*. ACM, 2010, pp. 333–342.

[53] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, and M. Hansen, “Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype,” in *Proc. of the 4th workshop on Embedded Networked Sensors*. ACM, 2007, pp. 13–17.

[54] J. Lee, A. Banerjee, and S. K. Gupta, “Mt-diet: Automated smartphone based diet assessment with infrared images,” in *Proc. of the International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2016, pp. 1–6.

[55] J. Zhang, B. Guo, H. Chen, Z. Yu, J. Tian, and A. Chin, “Public sense: Refined urban sensing and public facility management with crowdsourced data,” in *Proc. of the 7th International Symposium on Ubicom Frontiers - Innovative Research, Systems and Technologies (UFirst)*. IEEE, 2015, pp. 1407–1412.

[56] E. Koukoumidis, L.-S. Peh, and M. R. Martonosi, “Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory,” in *Proc. of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 2011, pp. 127–140.

[57] Q. Wang, B. Guo, G. Peng, G. Zhou, and Z. Yu, “Crowdwatch: pedestrian safety assistance with mobile crowd sensing,” in *Proc. of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp): Adjunct*. ACM, 2016, pp. 217–220.

[58] X. Zhang, H. Gong, Z. Xu, J. Tang, and B. Liu, “Jam eyes: a traffic jam awareness and observation system using mobile phones,” *International Journal of Distributed Sensor Networks*, vol. 2012, 2012.

[59] S. Morishita, S. Maenaka, D. Nagata, M. Tamai, K. Yasumoto, T. Fukukura, and K. Sato, “Sakurasensor: quasi-realtime cherry-lined roads detection through participatory video sensing by cars,” in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2015, pp. 695–705.

[60] P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, and M. Satyanarayanan, “Scalable crowd-sourcing of video from mobile devices,” in *Proc. of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 2013, pp. 139–152.

[61] B. Guo, Y. Liu, W. Wu, Z. Yu, and Q. Han, “Activecrowd: A framework for optimized multitask allocation in mobile crowdsensing systems,” *IEEE Trans. on Human-Machine Systems*, vol. 47, no. 3, 2017, pp. 392-403.

[62] H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. Barnes, “icrowd: Near-optimal task allocation for piggyback crowdsensing,” *IEEE Trans. on Mobile Computing*, vol. 15, pp. 2010–2022, 2016.

[63] D. Zhang, H. Xiong, L. Wang, and G. Chen, “CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint,” in *Proc. of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 703-714, 2014.

[64] S. Ji, Y. Zheng, and T. Li, “Urban sensing based on human mobility,” in *Proc. of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 1040-1051.

[65] B. Guo, H. Chen, Z. Yu, W. Nan, X. Xie, D. Zhang, and X. Zhou, “Taskme: toward a dynamic and quality-enhanced incentive mechanism

for mobile crowd sensing," *International Journal of Human-Computer Studies*, vol. 102, no. 6, 2017, pp. 14-26.

[66] R.K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32-39, 2011.

[67] L.G. Jaimes, J. Idalides, and A. Raij, "A survey of incentive techniques for mobile crowd sensing," *IEEE Internet of Things Journal*, vol. 2, no.5, pp. 370-380, 2015.

[68] M. Dong, X. Liu, Z. Qian, A. Liu, and T. Wang, "QoE-ensured price competition model for emerging mobile networks," *IEEE Wireless Communications*, vol. 22, no. 4, pp. 50-57, 2015.

[69] J. P. Rula and F. E. Bustamante, "Crowdsensing under (soft) control," in *Proc. of the 2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 2236–2244.

[70] R. Kawajiri, M. Shimosaka, and H. Kashima, "Steered crowdsensing: Incentive design towards quality-oriented place-centric crowd sensing," in *Proc. of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2014, pp. 691–701.

[71] X. Zhang, Z. Yang, Y. Liu, and S. Tang, "On Reliable Task Assignment for Spatial Crowdsourcing," *IEEE Transactions on Emerging Topics in Computing*, 2017 (to appear).

[72] Y. Liu, B. Guo, Y. Wang, W. Wu, Z. Yu, and D. Zhang, "TaskMe: multi-task allocation in mobile crowd sensing," in *Proc. of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 403-414.

[73] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 745-754.

[74] B. Guo, H. Chen, Q. Han, Z. Yu, D. Zhang, and Y. Wang, "Worker contributed data utility measurement for visual crowdsensing systems," *IEEE Trans. on Mobile Computing*, vol. 16, no. 8, 2017, pp. 2379-2391.

[75] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proc. of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)*. ACM, 2003, pp. 27–34.

[76] N.D. Lane, et al., "Piggyback CrowdSensing (PCS): energy efficient crowdsourcing of mobile sensor data by exploiting smartphone app opportunities," in *Proc. of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013.

[77] H. Xiong, D. Zhang, L. Wang, and H. Chaouchi, "EMC^3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," IEEE Transactions on Mobile Computing, vol. 14, no. 7, pp. 1355-1368, 2015.

[78] F. Xiao, Z. Jiang, X. Xie, L. Sun, and R. Wang, "An energy-efficient data transmission protocol for mobile crowd sensing," *Peer-to-Peer Networking and Applications*, vol. 10, no. 3, pp. 510-518, 2017.

[79] L. Chen, et al., "EnUp: Energy-Efficient Data Uploading for Mobile Crowd Sensing Applications," in *Proc. of IEEE UIC'16*, 2016.

[80] W. Sun and J. Liu, "Congestion-Aware Communication Paradigm for Sustainable Dense Mobile Crowdsensing," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 62-67, 2017.

[81] M. Dong, K. Ota, and A. Liu, "RMER: Reliable and energy-efficient data collection for large-scale wireless sensor networks," *IEEE Internet of Things Journal*, vol. 3, no. 4, pp. 511-519, 2016.

[82] D. Wu, Y. Zhang, J. Luo, et al., "Efficient data dissemination by crowdsensing in vehicular networks," in Proc. of IEEE IWQoS'14, 2014, pp. 314-319.

[83] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2005, pp. 886–893.

[84] Z. Wen, W. Zhu, J. Ouyang, P. Liu, Y. Du, M. Zhang, and J. Gao, "A robust and discriminative image perceptual hash algorithm," in *Proc. of the 4th International Conference on Genetic and Evolutionary Computing (ICGEC)*. IEEE, 2010, pp. 709–712.

[85] S. Thrun and J. J. Leonard, "Simultaneous localization and mapping," *Springer handbook of robotics*. Springer, 2008, pp. 871–889.

[86] J. Chow, "Multi-sensor integration for indoor 3d reconstruction," Ph.D. dissertation, University of Calgary, 2014.

[87] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz, "Scene reconstruction and visualization from community photo collections," *Proc. of the IEEE*, vol. 98, no. 8, pp. 1370–1390, 2010.

[88] C. Wu, "Visualsfm: A visual structure from motion system," Available: *http://ccwu.me/vsfm/* (accessed on 14 Jan. 2017).

[89] N. Snavely et al., "Bundler: Structure from motion (sfm) for unordered image collections," Available: *http://www.cs.cornell.edu/snavely/bundler/* (accessed on 14 Jan. 2017).

[90] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *Proc. of the 12th International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 72–79.

[91] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.

[92] R. Kwitt and A. Uhl, "Image similarity measurement by kullback-leibler divergences between complex wavelet subband statistics for texture retrieval," in *Proc. of the 15th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2008, pp. 933–936.

[93] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2963–2970.

[94] J. R. Parker, Algorithms for image processing and computer vision. John Wiley & Sons, 2010.

[95] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248-255.

[96] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958-1970, 2008.

[97] Y. Lin, F. Lv, S. Zhu, et al., "Large-scale image classification: fast feature extraction and SVM training," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1689-1696.

[98] O. Russakovsky, J. Deng, H. Su, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[99] Z. Akata, F. Perronnin, and Z. Harchaoui, et al., "Good practice in large-scale learning for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 507-520, 2014.

[100] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012.

[101] A. Karpathy, G. Toderici, S. Shetty, et al., "Large-scale video classification with convolutional neural networks," in *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725-1732.

[102] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," Computer Networks, vol. 51, no. 4, pp. 921-960, 2007

[103] F. G. Yap and H.H. Yen, "A survey on sensor coverage and visual data capturing/processing/transmission in wireless visual sensor networks," *Sensors*, vol. 14, no. 2, pp.3506-3527, 2014.

[104] V.P. Munishwar and N.B. Abu-Ghazaleh, "Coverage algorithms for visual sensor networks," ACM Transactions on Sensor Networks, vol. 9, no. 4, 2013.

[105] C. Ding, J.H. Bappy, J.A. Farrell, and A.K. Roy-Chowdhury, "Opportunistic Image Acquisition of Individual and Group Activities in a Distributed Camera Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 664-672, 2017.

[106] B. Guo, Z. Yu, X. Zhou, "A Data-Centric Framework for Cyber-Physical-Social Systems," *IEEE IT Professional*, vol. 17, no. 6, pp. 4-7, 2015.

[107] Y. Hu, X. Wang, X. Gan, "Critical sensing range for mobile heterogeneous camera sensor networks," in *Proc. of IEEE INFOCOM*, 2014, pp. 970-978.

[108] Y. Ding and L. Xiao, "SADV: Static-node-assisted adaptive data dissemination in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no.5, pp. 2445-2455, 2010.

[109] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30-39, 2017.

[110] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications," *IEEE Access*, 2017 (to appear).

[111] Y. C. Hu, M. Patel, D. Sabella, et al. "Mobile edge computing—A key technology towards 5G," *ETSI White Paper*, 2015.

[112] Y. Wu and G. Cao, “VideoMec: A Metadata-Enhanced Crowdsourcing System for Mobile Videos,” in *Proc. ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2017.

[113] Y. Wu, Y. Wang, and G. Cao, “Photo Crowdsourcing for Area Coverage in Resource Constrained Environments,” in *Proc. IEEE INFOCOM*, 2017.

[114] M. Satyanarayanan, P. Simoens, Y. Xiao, et al. “Edge analytics in the internet of things,” *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24-31, 2015.

[115] T. X. Tran, A. Hajisami, P. Pandey, et al., “Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges,” *IEEE Communications Magazine*, vol. 55, no. 4, pp. 54-61, 2017.

[116] M. Redi, D. Quercia, L. T. Graham, and S. D. Gosling, “Like partying? your face says it all. predicting the ambiance of places with profile pictures,” in *Proc. of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, 2015.

[117] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, “Anonysense: privacy-aware people-centric sensing,” in *Proc. of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 2008, pp. 211–224.

[118] D. Christin, “Privacy in mobile participatory sensing: current trends and future challenges,” *Journal of Systems and Software*, vol. 116, pp. 57–68, 2016.

[119] T. Winkler and B. Rinner, “Security and privacy protection in visual sensor networks: A survey,” *ACM Computing Surveys*, vol. 47, no. 1, 2014.

[120] A. Ker¨anen, J. Ott, and T. K¨arkk¨ainen, “The one simulator for dtn protocol evaluation,” in *Proc. of the 2nd International Conference on Simulation Tools and Techniques*. ICST, 2009, p. 55.

[121] L. Chen, D. Zhang, X. Ma, L. Wang, S. Li, Z. Wu, and G. Pan, “Container port performance measurement and comparison leveraging ship gps traces and maritime open data,” *IEEE Trans. on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1227–1242, 2016.

[122] D. Zhang, B. Guo, and Z. Yu, “The emergence of social and community intelligence,” *Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[123] D. Yang, D. Zhang, and B. Qu, “Participatory cultural mapping based on collective behavior data in location-based social networks,” *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, p. 30, 2016.

Bin Guo is currently a professor at the Northwestern Polytechnical University, China. He received his Ph.D. degree in computer science from Keio University, Japan in 2009. His research interests include ubiquitous computing, mobile crowd sensing, and HCI. He is a senior member of IEEE.

Qi Han is an associate professor of computer Science from Colorado School of Mines, US. She obtained her Ph.D. degree in Computer Science from the University of California-Irvine in August 2005. Her research interests include mobile crowd sensing and ubiquitous computing. She is a member of IEEE.

Huihui Chen is currently a Ph.D. candidate at Northwestern Polytechnical University. She received her M.S. degree from Zhengzhou University in 2006. Her research interests include ubiquitous computing and mobile crowd sensing.

Longfei Shangguan is currently a post-doc researcher at Princeton University. He received his Ph.D. degree from Hong Kong University of Science and Technology in 2015. His research interests include Internet of Things and mobile computing.

Zimu Zhou is currently a post-doc researcher at ETH Zurich. He received his Ph.D. degree from Hong Kong University of Science and Technology in 2015. His research interests include ubiquitous computing and mobile systems.

Zhiwen Yu is currently a professor at the Northwestern Polytechnical University, China. He has worked as an Alexander Von Humboldt Fellow at Mannheim University, Germany from Nov. 2009 to Oct. 2010. His research interests cover ubiquitous computing and HCI. He is a senior member of IEEE.