Survey of Visual Crowdsensing

Shuying Zhai¹,a, Ru Li¹, b

¹Xi’an Mingde Institute of Technology, Fenghe campus of Northwestern Polytechnical University, Xi’an 710124, China
aemail: Zhaisy@npumd.edu.cn , bemail: Liru@npumd.edu.cn

Abstract. In recent years, Visual Crowdsensing (VCS) that sensed through images and video, has become a predominant sensing paradigm of Mobile Crowdsensing (MCS), which is one of the current research hotspots. VCS requires people to capture the details of sensing objects in the real world in the form of pictures or video, which is widely used in various fields. However, there is no article summarizing the development and current situation of VCS. To this end, this paper summarized the latest applications of VCS, including floor plan generation, indoor scene reconstruction, outdoor scene reconstruction, event reconstruction, indoor localization, indoor navigation and disaster relief, and summarized some unique problems of VCS at present.

1. Introduction

With the development of computer technology and the expansion of the information world, cameras, microphones and other sensors are embedded in our daily portable mobile devices (smart phones, wearable devices). As a new way of sensing, Visual Crowd sensing is more and more well known. Visual Crowd sensing network uses existing mobile devices to complete sensing tasks, relying on cellular data, Wi-Fi, Bluetooth and other data transmission, without special deployment and maintenance, it can easily complete large-scale and fine-grained sensing. Through the intelligent computing of mobile group, users of mobile terminals can cooperate consciously or unconsciously to complete large-scale and high complexity social sensing tasks that are difficult for individuals to achieve. It has a wide application prospect in urban management, environmental monitoring, traffic conditions and so on.

Mobile Visual Crowdsensing can accomplish sensing tasks in many ways, such as numerical value (air quality), audio, picture / video. Among these ways of sensing, the way of using the built-in camera of mobile devices for sensing [1] has attracted more and more attention from academia and industry in recent years. In this context, Professor Guo Bin proposed the concept of Visual Crowd sensing, and made a comprehensive review of the task model, characteristics and key technologies of Visual Crowd sensing. Visual Crowd sensing requires users to obtain the detailed information of perceptual objects in the real world in the form of pictures or videos. Pictures and videos have attracted wide attention because they carry more abundant information. Previous studies, such as SISE [2], Firework [3], IONavi [4] and BEES [5], have proved the effectiveness of Visual Crowd sensing. In most cases, the Visual Crowdsensing is better than the traditional visual sensing which relies on fixed cameras for monitoring.

In Mobile Visual Crowdsensing, compared with other types of sensing, images and videos perceived by Visual Crowdsensing can contain more abundant information, which is more closely related to the environment (such as depth of field, shooting angle.), the volume of data items is larger, and the data processing is more complex. At the same time, Visual Crowd sensing also faces some unique problems, such as multi-dimensional coverage requirements, data redundancy identification and elimination.
far, Visual Crowd sensing has been widely used in various fields, but there is almost no domestic literature on the research status of Visual Crowd sensing. On the basis of literature [3], this paper classifies the applications of Crowdsensing, reviews the latest research on Visual Crowd sensing, and comprehensively summarizes the applications and challenges of Visual Crowd sensing, so as to provide reference for the further development of Visual Crowdsensing.

2. Application of Visual Crowd sensing

The existing applications of Visual Crowd sensing can be roughly divided into the following categories: plan generation, indoor scene reconstruction, outdoor scene reconstruction, event reconstruction, indoor positioning, indoor navigation, disaster rescue and urban sensing.

2.1. Building plan

Building plan is an important part of architectural design, which reflects the functional needs of the building, the plane layout and the composition of the plane, and shows the relationship between the physical characteristics of the room and space. It is very important for applications such as indoor positioning and indoor navigation. Previous indoor plan reconstruction systems rely heavily on inertial sensors, and the edge of the indoor scene is usually blocked by furniture and other objects, and most users cannot access the restricted indoor area, which makes the system unable to generate more accurate results. Images and videos can provide more abundant environmental information (geometry information, depth of field, color, etc.). Crowdmap [6] combines the temporal and spatial relationship between each successive frame of crowdsourcing video with inertial sensor to generate a more accurate indoor plan. Jigsaw proposed in reference [7] obtains the location, size and direction information of landmark objects in the image through Visual Crowd sensing, obtains the spatial relationship between adjacent landmark objects through inertial sensors, calculates the coordinates and orientation of objects, and then combines the user's moving trajectory with coordinates to generate a complete indoor plant.

2.2. Indoor scene reconstruction

Compared with the floor plan which only expresses the structure of buildings, scene reconstruction contains more information and plays an important role in fire rescue, augmented reality and indoor navigation. Traditional outdoor scene reconstruction technology mostly relies on special data acquisition equipment, and has complex and time-consuming offline processing mechanism, so it is difficult to be directly applied to indoor scene reconstruction, and this process can be easily and efficiently completed through Visual Crowd sensing. Reference [8] developed a smart phone application, which combines the video and image data captured by the mobile phone with accelerometer and gyroscope to visualize and reconstruct indoor scenes such as home and office, without any special equipment or training, and can also generate an accurate floor plan. To solve the problem that the existing indoor semantic plan will not be automatically updated with the environment, a mobile Crowdsourcing system SISE is proposed in reference [2]. In this paper, a data model called engraph is proposed to describe the indoor entities. The entities in the indoor plan are corresponding to those in the crowdsourcing image, and an algorithm is proposed to detect and locate the changed entities. Finally, in SISE, the indoor entity, its semantics and engraph are abstracted by the photos and inertial data submitted by users, so as to update the indoor semantic plan automatically.

2.3. Outdoor scene reconstruction

Outdoor scene reconstruction has important applications in Panoramic Map (such as Google Street View), augmented reality, urban monitoring and urban planning. In reference [9], an automatic storage system CrowdGIS is proposed. The information stored in the current map service is often out of date or even missing. CrowdGIS uses street view and crowdsourcing data to estimate the potential relationship between the location and the actual location through neural network analysis, so as to determine the user's shooting location and update the map data in the server. In reference [10], an online game PhotoCity is proposed, which enables players to take photos with specific angles (compared with the existing
incomplete 3D models in the database, and required to overlap or match them) through some incentive mechanisms, and then realizes 3D construction of 2D images through computer vision technology. In reference \[11\], a smart phone application crowd-pan-360, which can generate 360° panoramic images, is proposed. The system can sense the environment information, intelligently associate the user's position information with the information of various sensors (acceleration sensors, etc.), reduce the positioning error, effectively identify the user's environment, and enter the strange ring when the user enters. The panoramic image is generated at the scene time. Reference \[12\] proposed a set of key frame selection algorithm, which can generate large-scale outdoor panorama very quickly by using crowdsourced mobile video with geographical markers.

2.4. Indoor positioning
Because the mainstream positioning technology relies too much on wireless signals, two different locations may have the same RSS fingerprint in extremely complex cases, which leads to positioning errors.

In view of this situation, reference \[13\] proposed an image aided positioning system Argus based on mobile terminal. Argus extracts geometric constraints from crowdsourced image data, and maps these geometric constraints to RSS fingerprint space to distinguish the location information with similar fingerprints and reduce the fuzziness of fingerprints. Document \[14\] proposes a CSP framework. CSP collects image and audio data related to user's location through crowdsourcing. These data contain rich prompt information (including text prompt and physical object), so that CSP can abstract the physical location at a higher level according to the meaning of physical location to people, such as work, diet, exercise. Abstract and classify physical locations as logical locations.

2.5. Event reconstruction
With the popularity of smart phones and mobile Internet, people are more and more used to recording their lives in the form of pictures or videos, and sharing them through social networks. The analysis of these images and videos will help us to review the event after it happened. In reference \[15\], an on-demand system movi based on smart phones is proposed. It can sense the surrounding environment through smart phones and wait for the event trigger to wake it up (such as laughter). After the event, movi can splice the recordings and pictures on different mobile phones into a video clip, and its effect can be similar to that of the whole event edited by professional workers. Video clips are comparable. Reference \[16\] proposed an algorithm to locate events in the city in time and space by using image information on social networks (such as Instagram). It allows users to experience these events remotely with eyewitness lens without taking part in urban activities.

Each smart phone can only capture video and image in a limited perspective and distance, so the generated video is relatively single and the event information is not rich enough. However, in the same event, different participants can capture multiple video clips from different angles and distances. Reference \[17\] proposed a framework to capture the same spatiotemporal events and estimate their relative time offsets by analyzing the user generated video generated by sound matching. Finally, the user generated video from multiple unedited cameras is grouped and aligned to reconstruct the event. People can review the events from different perspectives through the videos uploaded by the event participants, but not every user can experience the event from their own satisfactory perspective. It may take a lot of time to find a satisfactory perspective in the uploaded videos. Reference \[18\] developed a query system to solve this problem. The system generates a lightweight 3D scene structure. The user can select a 2D region as the query condition in any related video, and the system will return the video related to the region to the user.

Traditional cloud centric real-time video analysis needs to load all video resources to a centralized cluster in advance, which has large response delay and high data transmission cost. In reference \[3\], a general computing framework firework is proposed, which makes the distributed data processing of IOE application more convenient through visual data view and service composition. Compared with traditional cloud centric solutions, firework greatly reduces latency and bandwidth costs. In reference
[19], a hybrid edge Cloud Architecture Based on android is proposed and demonstrated by a case. In the real scene of Portugal Volleyball League, the real-time video recorded by users is put into the edge 21 computer science cloud in 2019, which is synchronized and cached by raspberry pie, so that nearby users can share video without using any other network infrastructure, effectively reducing user delay and bandwidth occupation of network infrastructure. Reference [20] proposes a framework for video collection and integration, which is used to generate video clips containing participants' perspectives. The innovation of the framework is that it supports real-time transmission, processing and display of videos recorded by hundreds of users in ultra-dense Wi-Fi environment.

3. Problems of Visual Crowd sensing

The characteristics of Visual Crowd sensing bring it advantages as well as problems. Some problems still need to be solved are as follows.

3.1. Diversity of Equipment

Compared with the traditional mobile Crowdsensing task, Visual Crowd sensing usually requires the acquisition of sensing data from different angles and directions, and needs to consider the multi-dimensional scenario when task allocation, and the sensing nodes may be smart phones, cars connected to the network, wearable devices, etc., and the network access mode, stability, speed, acquisition and processing of different devices. There are great differences in the ability of perceptual information. In this case, it becomes more challenging to complete specific tasks within certain resource constraints.

3.2. Redundant filtering

Visual Crowd sensing task will collect a large number of original photo sets, but a considerable part of them are redundant in content or semantics (according to the task coverage definition [3]), and the redundant photos should be removed as much as possible when transmitting photos, Select a representative subset of photos for transmission, or upload the metadata of the photos for matching on the server side, determine the value and then perform the real transmission, so as to transmit more valuable visual information under limited bandwidth resources as much as possible.

3.3. Image matching

Visual Crowdsensing tasks often need to process a large number of crowdsourced images, and according to different task requirements, it needs to have image matching, three-dimensional modelling, image content recognition and image tagging capabilities, so lightweight and robust computer vision algorithms Is essential. Image matching is usually used to detect or group redundant information in visual information; 3D modelling can be used to generate indoor maps or 3D maps; image content recognition and image tagging help to group image information and extract the information Useful information such as words or symbols.

4. Conclusion

This paper introduces the Visual Crowd sensing, and focuses on the summary and comparison of the latest applications of Visual Crowd sensing, including eight aspects: plan generation, indoor scene reconstruction, outdoor scene reconstruction, event reconstruction, indoor positioning, indoor navigation, disaster rescue and urban sensing. Some problems of Visual Crowd sensing are also discussed. It includes the diversity of equipment, redundant image filtering, effective data transmission, image quality evaluation and image matching and processing. In the future work, we will further study the communication, processing and user privacy of visual data. We believe that VC's will be deployed in various types of applications in the next few years, which will naturally integrate into people's lives and provide people with more intelligent and humanized services.
Acknowledgments
This work was partially supported by the Scientific Research Plan of Ministry of Education of Shaanxi Province (No. 18JK1169) and the Scientific Research Foundation of Mingde Institute of Technology(2019XY01L01)

References
[1] WANG L, ZHANG D, WANG Y, et al. Sparse mobile Crowdsensing: challenges and opportunities[J]. IEEE Communications Magazine,2016,54(7):161-167.
[2] TENG X, GUO D, GUO Y, et al. SISE: Self-updating of Indoor Semantic Floorplans for General Entities[J]. IEEE Transactions on Mobile Computing,2018, PP (99):1.
[3] ZHANG Q, ZHANG Q, SHI W, et al. Firework: Data Processing and Sharing for Hybrid Cloud-Edge Analytics[J]. IEEE Transactions on Parallel & Distributed Systems,2018, PP(99):1.
[4] TENG X, GUO D, GUO Y, et al. IONavi: an indoor-outdoor navigation service via mobile crowdsensing[J]. ACM Transactions on Sensor Networks,2017,13(2):12.
[5] ZUO P, HUA Y, LIU X, et al. BEES: Bandwidth-and Energy-Efficient Image Sharing for Real-Time Situation Awareness[C]//IEEE, International Conference on Distributed Computing Systems. IEEE,2017:1510-1520.
[6] CHEN S, LI M, REN K, et al. Crowd Map: Accurate Reconstruction of Indoor Floor Plans from Crowdsourced Sensor-Rich Videos[C]//2015 IEEE 35th International Conference on Distributed Computing Systems. IEEE,2015:1-10.
[7] GAO R, ZHAO M, YE T, et al. Jigsaw: indoor floor plan reconstruction via mobile crowdsensing[C]//International Conference on Mobile Computing and NETWORKING. ACM,2014:249-260.
[8] SANKAR A, SEITZ S. Capturing indoor scenes with smartphones[C]//ACM Symposium on User Interface Software and Technology. ACM,2012:403-412.
[9] PENG Z, GAO S, XIAO B, et al. CrowdGIS: Updating Digital Maps via Mobile Crowdsensing[J]. IEEE Transactions on Automation Science &. Engineering,2017, PP (99):1-12.
[10] TUITE K, SNAVELY N, HSIAO D Y, et al. PhotoCity: training experts at large-scale image acquisition through a competitive game[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM,2011:1383-1392.
[11] RAYCHOU DHURY V, SHRIVASTAV S, SANDHA S S, et al. CROWD-PAN-360: Crowdsourcing Based Context-Aware Panoramic Map Generation for Smartphone Users[J]. IEEE Transactions on Parallel &. Distributed Systems,2015,26(8):2208-2219.
[12] KIM S H, LU Y, SHI J, et al. Key Frame Selection Algorithms for Automatic Generation of Panoramic Images from Crowdsourced Geo-tagged Videos[C]//International Symposium on Web and Wireless Geographical Information Systems. Springer, Berlin, Heidelberg, 2014:67-84.
[13] XU H, YANG Z, ZHOU Z, et al. Enhancing Wi-Fi-based localization with visual clues[C]//Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM,2015:963-974.
[14] CHON Y, LANE N D, LI F, et al. Automatically characterizing places with opportunistic crowdsensing using smartphones[C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012:481-490.
[15] BAO X, CHOUDHURY R R. MoVi: mobile phone based video highlights via collaborative sensing [C]//International Conference on Mobile Systems, Applications, and Services. ACM,2010:357-370.
[16] GIRIDHAR P, WANG S, ABDELZAHER T, et al. On localizing urban events with Instagram[C]//INFOCOM 2017-IEEE Conference on Computer Communications. IEEE, 2017:1-9.
[17] BANO S, CA VALLARO A. Discovery and organization of multi-camera user-generated videos of the same event [J. Information Sciences, 2015,302:108-121.
[18] CARLIER A, CALVET L, GURDJOS P, et al. Querying Multiple Simultaneous Video Streams with 3D Interest Maps [M]//Visual Content Indexing and Retrieval with Psycho-Visual Models. Springer, Cham, 2017:125-144.

[19] RODRIGUES J, MARQUES E R B, SILVA J, et al. Video Dissemination in Untethered Edge-Clouds: A Case Study [C]//IFIP International Conference on Distributed Applications and Interoperable Systems. Springer, Cham, 2018:137-152.

[20] BOHEZ S, DANEELS G, VAN HERZEELE L, et al. The crowd as a cameraman: on-stage display of crowdsourced mobile video at large-scale events [J]. Multimedia Tools and Applications, 2018, 77(1):597-629.