Research Statement

Bin Zhu
School of Computing and Information Systems, Singapore Management University
Tel: 6828 0270; Email: binzhu@smu.edu.sg
1 (Day) 12 (Month) 2023 (Year)

Background

Multimedia data such as text, audio, image and video are ubiquitous in our daily lives. Being able to understand these multimedia contents, undoubtedly, will benefit and even shape people's daily life in different aspects, ranging from information retrieval, entertainment, and social participation to health management. My research interest lies in **Human-Centered Multimedia Analysis**. Specifically, the objective is to conduct frontier research and develop cutting-edge technologies for processing, modeling, analyzing, and understanding multimedia contents, that facilitate natural and immersive human experience and exert positive impact for our society.

Toward this target, I have contributed to the acquisition of machine intelligence from raw multimedia inputs, such as reasoning object relations in egocentric videos, learning latent cross-modal representation for multimedia search, generating images from natural language descriptions, and exploring food computing for health management. In the following, I summarize my research contributions from three perspectives and then introduce my future research agenda.

Research Areas

Multimedia search and creation

My research focuses on advancing the field of multimedia search and creation, with a particular emphasis on developing interpretable and generalizable cross-modal retrieval system (e.g., text-to-image retrieval), as well as manipulable cross-modal generation system (e.g., text-to-image synthesis). The challenge for multimedia search and creation is the heterogeneous nature of multimedia data. As a result, how to develop algorithms to learn cross-modal feature representation, how to explain the search results beyond similarity measurement for the users, how to bridge the gap between modalities for multimedia creation, as well as how to enable model to be robust and effective with domain shift, are worth for exploration.

To bridge the gap among different modalities for multimedia data, the common solution is to learn a shared latent space for similarity measurement. Nevertheless, the model training for multimedia search is a black box, thus lacking interpretability. We address the issue by unifying the retrieval and generation. Specifically, Recipe Retrieval Generative Adversarial Network (R2GAN) [1] is proposed to learn compatible cross-modal features in an adversarial way, and generate images from recipes to explain the search results. To achieve this, R2GAN is particularly designed to have two discriminators, with one to distinguish between real and fake images as in common practice, and the other to predict whether a fake image is generated from image or recipe embedding. Furthermore, to generate photo-realistic image from recipe, causality-based text-to-image synthesis [2] is explored by simulating the cooking process in a step-to-step manner. The proposed CookGAN is able to mimic visual effect in causality chain and progressively upsample image, while preserving fine-grained details and showing semantic manipulability. In addition, we also explore cross-modal retrieval with domain shift where the multimedia data could be collected from different regions [3] or even the text data is in different languages [5], as well as generating caption from image in an unsupervised manner [12].

Egocentric video understanding

My research focuses on leveraging egocentric video understanding to enhance the immersive experiences of virtual and augmented reality systems. Captured from wearable devices such as GoPro and Google glasses, egocentric videos provide a first-person view that aligns with the users' perceptions and interactions in virtual environments. As a result, Egocentric video understanding is significantly promising for metaverse. Nevertheless, the current research on video understanding primarily focuses on the third-person point of view [8-10], since it's more convenient for an independent spectator to record videos and the datasets are much richer. As egocentric videos require special wearable devices to record, building egocentric video datasets and benchmarks are non-trivial. In contrast to third-person videos, egocentric videos not only vary in view of recording, but also camera movement, fine-grained human activities as well as interaction between hands and objects. Egocentric video understanding is a promising but challenging research direction. To address the above issues, I have already made significant contributions in this direction by developing new algorithms [7], establishing large-scale dataset and benchmarks [6] (https://epic-kitchens.github.io/VISOR/), as well as organizing challenges (https://epic-kitchens.github.io/2023).

Egocentric video action recognition in different domains: To address the challenge of domain shift in egocentric videos collected at different time, geographical locations, and environments, we propose a novel multi-modal adversarial learning method for unsupervised video domain adaptation [7]. Our approach incorporates the temporal information of videos in different domains, as well as the diversity of adaptability between modalities, such as RGB and optical flow. Additionally, we are currently exploring audio as an extra modality to learn more transferable and discriminative video representation.

Large-scale egocentric dataset and benchmarks: High-quality egocentric video datasets are the cornerstone and fuel for egocentric research. We introduce EPIC-KITCHENS VISOR (VIdeo Segmentations and Object Relations) [6], a new dataset of pixel annotations and a benchmark suite for segmenting hands and active objects in egocentric video. In particular, the pixel-level annotations are maintained with short-term and long-term consistency even with object transformation, e.g., an onion is peeled, diced and cooked in one video. In total, we publicly release 272K manual semantic masks of 257 object classes, 9.9M interpolated dense masks, 67K hand-object relations, covering 36 hours of 179 untrimmed videos. Three benchmarks are introduced with the annotations: (1) Semi-Supervised Video Object Segmentation (VOS), aims to track the object segments across consecutive actions. (2) Hand Object Segmentation (HOS), is to predict the contact between hands and the active objects, as well as both the relation and accurate segmentations of hands and objects. (3) Where Did This Come From (WDTCF)?, traces back a given segment to the container it came from (e.g., milk from the fridge, and a plate from a particular cupboard). I was deeply involved in the entire process of the project, from video data preparation and processing, annotation communication, quality control, benchmark formulation, dataset revealing to challenge organization. In particular, I am in charge of the third benchmark WDTCF focusing on long-term object relation reasoning.

EPIC-KITCHENS-100 Challenges: To further boost the development of egocentric video understanding and benefit the research community, I serve as one of the organizing committee members for the EPIC-KITCHENS-100 Challenges in 2022 and 2023. EPIC-KITCHENS-100 Challenges consist of nine challenges, covering a wide range of tasks, such as action recognition, action detection, hand-object segmentation, and video object segmentation. I am responsible for unsupervised domain adaptation for action recognition. The task aims to assign a (verb, noun) label to a trimmed segment, following the Unsupervised Domain Adaptation paradigm: a labeled source domain is used for training, and the model needs to adapt to an unlabeled target domain. The winners are announced at 10th EPIC@CVPR2022 workshop. EPIC-KITCHENS-100 2023 Challenges are ongoing and the winners will be announced in CVPR2023 workshop as well.

Food computing for wellness intelligence

My research focuses on the development of food computing techniques to establish the relationship between food and health, with a particular emphasis on food recognition, ingredient recognition and cross-modal recipe retrieval [1, 4, 11]. As the famous quote, "You are what you eat", food has a significant impact on every aspect of human life. By recognizing the food categories, analyzing the dish composition, logging the meal consumption, as well as estimating the calorie and nutrition information, we can assist people in managing food intake and avoid diseases related to diet, e.g., obesity, hypertension, diabetes and cardiovascular diseases. Nevertheless, recognizing the food categories and ingredients is challenging due to complex composition of dishes and variation of ingredients with different cooking and cutting methods.

To support this research, we introduce Vireo Food-251[11], a large-scale Chinese food dataset of 169,673 images containing 251 popular Chinese food and 406 ingredients. We propose multi-task learning and region-wise recognition approaches for food category and ingredient recognition. By predicting the food category and ingredient composition from dish images, we can estimate calorie and nutrition information for automatic food logging. Long-term records of food intake can be used for health trend prediction and personalized diet suggestions in collaboration with nutritionists. Through my research, I aim to develop a comprehensive health management system that leverages food computing techniques for wellness intelligence. This system could provide individuals with a better understanding of their diet and its impact on their overall health, enabling them to make informed decisions about their food choices and ultimately lead a healthier lifestyle.

Future research agenda

In summary, I have devoted my efforts to developing intelligent multimedia techniques and systems to improve the human experience. In the future, I am excited to advance Human-Centered Multimedia Analysis by exploring the following research questions:

- **Long-form egocentric video understanding**. Egocentric untrimmed videos are usually long in length, ranging from several minutes to hours, and contain rich fine-grained activities (e.g., a cooking video could contain hundreds of actions, such as peeling, cutting, slicing, washing, chopping and frying, etc.). Most of works focus on understanding trimmed short-term video clips (e.g., 10 seconds). It would be helpful to know what is happening in the video clip, such as recognizing the human actions within the short clips. Nevertheless, it would be difficult to temporally reason why the action happens without the context in history. For instance, to answer the questions like 'Where did I put my key?', it is necessary to analyze the long videos to obtain contextual information in the past. I would like to investigate long-form egocentric video understanding by leveraging the capability of large language models (e.g., GPT) and lightweight modalities (e.g., audio) to capture the context information in the long term.

- **Personalized and controllable multimedia content creation**. Most of the current generative models still lack of personalized and controllable abilities. My future research will focus on addressing this issue by developing techniques and user-friendly computer systems for multimedia content creation. Specifically, I plan to investigate novel personalized and controllable approaches for text-to-image and text-to-video generation, that take into account user preferences by using prompts and various visual signals.

- **Intelligent health management system**. Despite great progress in food recognition and cross-modal recipe retrieval has been achieved, quantifying the dish portion, health trend prediction as well as engaging users in active logging are still open problems. I have started to investigate the problems of weight trend prediction and diet suggestion generation based on daily food consumption. I plan to further explore interactive ways for active user logging, develop new algorithms for health trend prediction, as well as build health management system. In addition, I plan to establish an egocentric video dataset and several benchmarks for food. The video recording will cover both food preparation, along with benchmarks such as ingredient recognition, portion size estimation, nutrition estimation etc.

Selected Publications and Outputs

[1] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. "R2GAN: Cross-modal recipe retrieval with generative adversarial network." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[2] Bin Zhu, and Chong-Wah Ngo. "CookGAN: Causality based text-to-image synthesis." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[3] Bin Zhu, Chong-Wah Ngo, and Jingjing Chen. "Cross-domain cross-modal food transfer." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 3762-3770. 2020.
[4] Bin Zhu, Chong-Wah Ngo, and Wing-Kwong Chan. "Learning from Web Recipe-image Pairs for Food Recognition: Problem, Baselines and Performance." IEEE Transactions on Multimedia 24 (2021): 1175-1185.

[5] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Wing-Kwong Chan. “Cross-lingual Adaptation for Recipe Retrieval with Mixup”. In Proceedings of the 2022 International Conference on Multimedia Retrieval.
[6] Ahmad Darkhalil*, Dandan Shan*, Bin Zhu*, Jian Ma*, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. "EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations." In 36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2022. (*co-first authors)
[7] Yuehao Yin, Bin Zhu, Jingjing Chen, Lechao Cheng, and Yu-Gang Jiang. "Mix-DANN and Dynamic-Modal-Distillation for Video Domain Adaptation." In Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[8] Yanbin Hao, Zi-Niu Liu, Hao Zhang, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. "Person-level action recognition in complex events via tsd-tsm networks." In Proceedings of the 28th ACM International Conference on Multimedia (grand challenge). 2020.
[9] Yanbin Hao, Chong-Wah Ngo, and Bin Zhu. "Learning to match anchor-target video pairs with dual attentional holographic networks." IEEE Transactions on Image Processing 30 (2021): 8130-8143.
[10] Yanbin Hao, Jingru Duan, Hao Zhang, Bin Zhu, Pengyuan Zhou, and Xiangnan He. "Unsupervised Video Hashing with Multi-granularity Contextualization and Multi-structure Preservation." In Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[11] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. "A study of multi-task and region-wise deep learning for food ingredient recognition." IEEE Transactions on Image Processing 30 (2020): 1514-1526.
[12] Yu, Jiarui, Haoran Li, Yanbin Hao, Bin Zhu, Tong Xu, and Xiangnan He. "CgT-GAN: CLIP-guided Text GAN for Image Captioning." In Proceedings of the 31st ACM International Conference on Multimedia. 2023.