Research Statement

Chong-Wah Ngo
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6828-4818; Email: cwngo@smu.edu.sg
15-Dec-2021

Background

Image, video and audio are now prevalent media of online communication. Machine understanding of multi-modal signals becomes essential to make sense of everyday activities, social and emotional signals. My research objective is to enable machine understanding of multimedia content, structure, semantics and the associate values, which can drive a variety of domains, ranging from multimedia search, surveillance event analysis, behavior understanding to wellness applications. My specific research interest is to seek new approaches to bridge the semantic gap between low-level audio-visual features and high-level human knowledge for building interactive and intelligent systems at the intersection of different media analysis.

We have made a tremendous progress on recognizing complex multimedia content in the space and time axes. However, these advancements are bottlenecked by the requirement of large, clean and balance data for model training. Furthermore, the trained models often cannot be directly deployed for real-world applications due to “black box operations” that hinder the downstream inferencing task. Building machine with explainable ability remains an open issue. Search is a basic function of various applications and is expected to have intelligence to infer user expectation and explain result. In multimedia search, queries are mostly ad-hoc written in natural language and the out-of-vocabulary (OOV) problem persists. Continuous expansion of visual vocabulary to deal with OOV is difficult to realize in practice. Even if this is possible, the same semantics might be interpreted differently in different context, which is not possible to tackle alone by relying on content to make sense of user information need. Therefore, how to leverage user search behavior to understand queries under the constraint of limited vocabulary size becomes an interesting research problem. User interacts with search results through a stream of operations between browsing and query refinement, which provides context for real-time understanding of queries and user expectation. My research aims to seek new solutions towards the integration of modern machine learning algorithms, user behavior understanding and knowledge graph inference to enable multi-faceted understanding of multimedia content with reasoning capability.

Among the various applications relevant to multimedia analysis, lifestyle behavior analysis, such as food computing for wellness, is a promising research direction. The topic is multi-disciplinary, involving content analysis, behavior science and healthcare, and has high potential impact to several burning issues such as chronic diseases prevention and self-management. My research seeks for recognition of daily signals (food, activity, emotion) and causality inference of personal historical data, aiming to enable risk analysis, lifestyle recommendation and behavior nudging using short-term and long-term observational data. There are numerous challenges towards this goal, such as the development of human-centric technologies to engage users in logging lifestyle data, robust cause-and-effect modeling of time-series data, and provision of human-like communication to nudge users for healthy lifestyle.

Research Areas

I have devoted my research career to advance understanding of multimedia data. My current interests range from the fundamental research in multimedia processing and understanding to the applied research in multimedia search, activity recognition and food computing.

Multimedia Search

Multimedia computing is challenging due to the mismatch between visual perception and high-level semantics. Developing large-scale semantic concept detectors, as "semantic filters for multimedia content", is regarded as a powerful way to bridge the semantic gap. I have conducted theoretical and empirical research in this direction – the so-called concept-based search paradigm. The success is limited by the size of concept bank and the problem of query drifting. Specifically, due to the lack of query context understanding, concepts are erroneously selected and improperly weighted. Consequently, the search generally suffers from high recall and low precision. With the rapid advance in deep learning, cross-modal embedding replaces concept-based search as the mainstream approach. The key idea is to project text queries and videos into a joint latent space for representation learning. Indexing is concept-free because videos are represented in high dimensional continuous space and indexed by the embeddings extracted from deep neural networks. Nevertheless, concept-free search suffers from black-box training and the learned embeddings are not interpretable. This creates the issue of search robustness.

My recent research has addressed this problem by proposing a dual-task model to interpret the semantics underlying the query and video embeddings. Specifically, the network is an encoder-decoder architecture, which learns cross-modal representation while decoding the representation into a list of concepts as semantic interpretation. The deep network is trained with video-caption pairs. However, a caption usually only describes a specific aspect of video content. Hence, the problem of missing labels is prevalent in cross-modal representation learning. Similarly, as query is formulated in free-form style, the same information need can be expressed using different words. We address this problem by proposing likelihood training [4] to encourage the decoder to interpret the semantics beyond the words observed in the training data. However, likelihood training suffers from the same issue as the maximum likelihood estimation (MLE). The decoder tends to predict frequent and co-occurred concepts resulting in semantic inconsistency in embedding interpretation. We address this problem by proposing unlikelihood training that adaptively supresses conflicting concepts based on video or query context. The two counterbalancing training strategies are combined to introduce perturbations in representation learning, aiming to improving consistency in interpretation and robustness in search. These works have seamlessly integrated concept-based and concept-free paradigms into a hybrid search. The retrieval can be performed in real-time, and the result is less sensitive to query expression and quality of training data. The network outperforms several state-of-the-art models, such as dual-coding, SEA and hybrid space. Our system is ranked at top-1 in the TRECVid evaluation campaign 2021, being commented as the only system that can deliver high retrieval performance while providing unique target videos not found by any other systems [2].

My current research also drives towards bridging the gap between data, system and human. Quality search results depend on the nature of a dataset (e.g., video domain, feature distribution), how a query is formulated, and the feedback is provided. Without a system that can understand the search-ability of a query, user can end up frustrating with either result mixed of true and false positives or “no hit” (aka null query problem). In interactive search, query formulation becomes an experience of “trial-n-error” when no clue is hinted of why a video is deemed as relevant [3]. With the progress that we have made for interpretable embedding, we are able to explore reinforcement learning algorithms to incorporate user feedback in the search loop [1]. Specifically, by episode sampling of user navigation paths, our system learns to seek a path that can maximize reward based on the continuous user feedback. During interactive search, a user can provide feedback, such as “this is not a red rock mountain” or “the man should hold microphone”. The feedbacks will be encoded together with the browsing and query history, and the policy network will plan for navigation path to select clips for user to inspect. This two-way system-user interaction is expected to reduce user burden from mental tiredness due to trial-n-error querying and exhaustive browsing.

Video Analytics

Cameras are everywhere, however, providing insights beyond moving pixels such as detecting specific activities from videos remains difficult. My research [5-6] focuses on activity recognition, aiming to develop robust and efficient representation learning techniques. The relevant works include researching effective spatio-temporal analysis of videos, localization of activity instances from full-length videos in real-time, online detection of the presence of instances within a predefined duration from the time when the activity begins. These research works aim to automate content understanding and large-scale video indexing, and the management of surveillance camera content for smart city development.

Recently, my work seeks for low-cost high-performance video representation learning by leveraging the strength of transformer in global attention modeling and CNN in local modeling. We have proposed TokShift Transformer, a novel zero-parameter, zero-FLOPs operator, for modeling long-range temporal relations. Specifically, the TokShift barely temporally shifts partial [Class] token features back-and-forth across adjacent frames. Then, we densely plug the module into each encoder of a plain 2D vision transformer for learning 3D video representation. It is worth mentioning that TokShift is a pure convolutional-free video transformer pilot for video understanding. We are exploring different ways to reduce the model complexity and computational cost of TokShift. We envision a hybrid neural network that is more powerful than the existing de-facto 3D-CNN in terms of capacity in modelling information dynamics, while is more lightweight than the transformer and CNN in terms of the number of network parameters and GFLOPs.

Food Computing and Lifestyle Behavior Tracking

Understanding the food content (e.g., ingredient, cooking method, nutrition, taste, smell) has been a hot research topic. The problem has potential to generate high impact to healthcare, nutrition and social science. I address this problem from the angle of multimedia, by analyzing multi-modal signals including fine-grained visual cues from images and procedural text descriptions from recipes, for cross-modal image-to-recipe retrieval [7-8]. Particularly, my research tackles the scalability of food recognition, for example, ingredient recognition for zero-shot recipe retrieval and ingredient localization for cross-modal recipe retrieval. These works contribute to joint food attributes recognition (ingredient, cooking methods), feature embedding learning between images and recipes, and synthesis of food images for explainable recipe retrieval.

Scaling up recognition to increase food coverage and engaging users in active food logging are two extreme challenges in food recognition. As process of food logging is complicated and not implicit as step counting, providing timely incentives to motivate users become essentially important. In short-term, my research drives towards transfer learning for food recognition and video-based food portion estimation. Domain shift refers to variations in food appearance and preparation due to cuisines and changes in geography regions, seasons or personal preferences. The motivation of transfer learning is to research small sample learning algorithms that can self-adapt a learnt model to address subtle variations due to domain changes across different cuisines. Current practice in photo-based food logging is by uploading picture to server for recognition. Food portion size estimation, which requires multiple pictures for 3D estimation, is difficult to realize due to limited network bandwidth. With the arrival of 5G network, it is possible to capture a video of food for recognition and provide AR/VR interface in real-time to guide users to quickly complete the logging process. Studying of the algorithms for interactive-based 3D estimation of portion size is of both research and practical values. The problem is significant because logging portion size is not intuitive to users, and even challenging to dietitians, due to the requirement to use different measurement units to quantize various food types. In long-term, my research will combine behavior science and nutrition science for self-motivated way of logging and extend beyond food to other lifestyle data including activity and emotion logs. The current logging method is considered “passive” for requiring users to explicitly input data. Chatbot, which can provide incentive by allowing users to enquire lifestyle relevant questions while logging, is envisioned to be a new interface for self-motivated logging. My research aims to investigate new approaches that users can log data by free-form question-answering, while the system can inference from past data, personal profile and medicine knowledge graph to intervene and nudge users in human-like manner.

Selected Publications and Outputs

1. Z. Ma, J. Wu, Z. Hou, C. W. Ngo, “Reinforcement Learning based Interactive Video Search,” *Multimedia Modeling*, 2022.
2. J. Wu, Z. Hou, Z. Ma, C. W. Ngo, “VIREO@TRECVid 2021 Ad Hoc Video Search,” *TRECVid Workshop*, 2021.
3. P. A. Nguyen, C. W. Ngo, “Interactive Search vs. Automatic Search: An Extensive Study on Video Retrieval,” *ACM Trans. on Multimedia Computing, Communications, and Applications*, June 2021.
4. J. Wu, C. W. Ngo, “Interpretable Embedding for Ad-hoc Video Search,” *ACM Multimedia*, 2020.
5. H. Zhang, Y. Hao, C. W. Ngo, “Token Shift Transformer for Video Classification,” *ACM Multimedia*, 2021.
6. Z. Qiu, T. Yao, C. W. Ngo, T. Mei, “Optimization Planning for 3D ConvNets,” *Int. Conf. on Machine Learning*, 2021.

7. H. T. Nguyen, C. W. Ngo, “Terrace based Food Counting and Segmentation,” *AAAI*, 2021.
8. J. J. Chen, B. Zhu, C. W. Ngo, T. S. Chua, Y. G. Jiang, “A Case Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition,” *IEEE Trans. on Image Processing*, 2021.