# M4MM'22: 1st International Workshop on Methodologies for Multimedia

Xavier Alameda-Pineda
Inria
Grenoble, France
xavier.alameda-pineda@inria.fr

Qin Jin
Renmin University of China
Beijing, China
qjin@ruc.edu.cn

Vincent Oria
New Jersey Institute of Technology
Newark, New Jersey, USA
oria@njit.edu

Laura Toni
University College London
London, United Kingdom
l.toni@ucl.ac.uk

## ABSTRACT

Transversely to all multimedia research, the methods and tools are often shared by several MM topics and applications. M4MM aims to foster discussion around fundamental methodologies and tools that are used in various multimedia research topics. To that aim, the technical program of the workshop consists of two keynote speakers, on complementary and very relevant methods and tools for MM, as well as four technical papers. The complete M4MM'22 workshop proceedings are available at:
https://dl.acm.org/doi/proceedings/10.1145/3552487

## CCS CONCEPTS

• **Computing methodologies**; • **Information systems** → **Multimedia information systems**; • **General and reference** → **Cross-computing tools and techniques**;

## KEYWORDS

Methods, tools, best practices, Multimedia

**ACM Reference Format:**
Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni. 2022. M4MM'22: 1st International Workshop on Methodologies for Multimedia. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3503161.3554769

## 1 INTRODUCTION

Methodologies and tools are critical foundations for developing multimedia systems and applications. M4MM: Methodology and Tools for Multimedia, is a forum for researchers and practitioners to discuss methodologies and tools developed for multimedia systems and applications. M4MM is interested in papers describing best practices, methodologies, and tools developed for general purpose multimedia applications or within specific multimedia application areas. For this first edition, we have selected 4 papers and invited a keynote. The topics discussed in the selected papers for this first edition are: Playing Lottery Tickets in Style Transfer Models by Meihao Kong et al., Optimal Tensor Bipartite Graph Learning by Yang Haizhou et al., HPFL: Federated Learning by Fusing Multiple Sensor Modalities with Heterogeneous Privacy Sensitivity Levels by Yuanjie Chen et al., and Boosting few-shot learning by self-calibration in feature space by Kaipeng Zheng et al.

## 2 INVITED SPEAKERS

The workshop program features two very relevant and complementary topics in multimedia. On the one side, Prof. Hayley Hung will discuss how to design and implement "tools for collecting, synchronising and annotating ecologically valid social behaviour in the wild." On the other side, Prof. Simon Leglaive will present recent progress on "audio-visual representation learning using a multimodal dynamical variational autoencoder."

Hayley Hung is an Associate Professor at Delft University of Technology, The Netherlands. She leads the Socially Perceptive Computing Lab. Her research focuses on devising novel Pattern Recognition and Machine Learning methods to automatically interpret group social and affective behavior during face-to-face human interactions. In 2015, she was awarded the Dutch Research Foundation (NWO) Career Talent Award for experienced researchers (Vidi). Her research contributions have also been recognized via an invited talk at the ACM (Association of Computing Machinery) Multimedia conference Rising Star Session (2016). She was nominated for outstanding paper at ACM International Conference Multimodal Interaction (ICMI) in 2011. She has been an active member of the research community in shaping interdisciplinary dialogue at the international level (e.g., panel at ACM Conference on Multimedia 2013,2019,2021 & ACM ICMI 2019, Lorentz Workshop, at Interdisciplinary Network of Group Research (INGroup) as a panelist in 2015, and workshop organizer in 2019). She contributes regularly to community organizational roles, notably Program Co-chair of Association of Computing Machinery Multimedia Conference (2019) and the International Conference on Affective Computing and Intelligent Interaction (2017). She is General Co-Chair of the ACM International Conference on Multimodal Interaction (2024). She is an Associate Editor of the IEEE Transactions on Affective Computing.

Simon Leglaive is a tenured Assistant Professor at CentraleSupélec and a researcher in the AIMAC team of the IETR laboratory,


*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3554769

a CNRS joint research unit in Rennes, France. He received the Engineering degree from Télécom Paris (Paris, France) and the M.Sc. degree in acoustics, signal processing and computer science applied to music (ATIAM) from Sorbonne University (Paris, France) in 2014. He received the Ph.D. degree from Télécom Paris in the field of audio signal processing in 2017. He was then a post-doctoral researcher at Inria Grenoble Rhône-Alpes, in the Perception team. His research focuses on audio signal processing and machine learning. He is mainly interested in Bayesian approaches for problems that consist in estimating latent signals of interest from noisy and/or incomplete observations. His recent work focuses on weakly-supervised methods with deep-learning-based generative models, essentially dynamical variational autoencoders.

# 3 CONTRIBUTED PAPERS

## 3.1 Playing Lottery Tickets in Style Transfer Models

Style transfer has achieved great success and attracted a wide range of attention from both academic and industrial communities due to its flexible application scenarios. However, the dependence on a pretty large VGG-based autoencoder leads to existing style transfer models having high parameter complexities, which limits their applications on resource-constrained devices. Compared with many other tasks, the compression of style transfer models has been less explored. Recently, the lottery ticket hypothesis (LTH) has shown great potential in finding extremely sparse matching subnetworks which can achieve on par or even better performance than the original full networks when trained in isolation. In this work, we for the first time perform an empirical study to verify whether such trainable matching subnetworks also exist in style transfer models. Specifically, we take two most popular style transfer models, i.e., AdaIN and SANet, as the main testbeds, which represent global and local transformation-based style transfer methods respectively. We carry out extensive experiments and comprehensive analysis and draw the following conclusions. (1) Compared with fixing the VGG encoder, style transfer models can benefit more from training the whole network together. (2) Using iterative magnitude pruning, we find the matching subnetworks at 89.2% sparsity in AdaIN and 73.7% sparsity in SANet, which demonstrates that Style transfer models can play lottery tickets too. (3) The feature transformation module should also be pruned to obtain a much sparser model without affecting the existence and quality of the matching subnetworks. (4) Besides AdaIN and SANet, other models such as LST, MANet, AdaAttN and MCCNet can also play lottery tickets, which shows that LTH can be generalized to various style transfer models.

## 3.2 Optimal Tensor Bipartite Graph Learning

This paper proposes an efficient multi-view clustering method called Optimal Tensor Bipartite Graph Learning for multi-view clustering (OTBGL). The model is a novel tensorized bipartite graph based multi-view clustering method with low tensor-rank constraint. Firstly, to substantially reduce the computational complexity, the bipartite graphs are leveraged from different views instead of the full similarity graphs of the corresponding views. Secondly, the similarity between bipartite graphs of different views is computed by minimizing the tensor Schatten p-norm as a tighter tensor rank approximation. In addition, the L1,2-norm of learned graphs are minimized to explore the spatial low-rank structures embedded in the intra-view graphs. Thirdly, an efficient algorithm suitable for processing large-scale data proposed. Extensive experimental results on six benchmark datasets indicate that the proposed OTBGL is superior to the state-of-the-art methods.

## 3.3 HPFL: Federated Learning by Fusing Multiple Sensor Modalities with Heterogeneous Privacy Sensitivity Levels

Solving classification problems to understand multi-modality sensor data has become popular, but rich-media sensors, e.g., RGB cameras and microphones, are privacy-invasive. Though existing Federated Learning (FL) algorithms allow clients to keep their sensor data private, they suffer from degraded performance, particularly lower classification accuracy and longer training time, than centralized learning. A Heterogeneous Privacy Federated Learning (HPFL) paradigm is proposed to capitalize on the information in the privacy insensitive data (such as mmWave point clouds) while keeping the privacy sensitive data (such as RGB images) private because sensor data are of diverse sensitivity levels. This is the first time multiple media/modalities with diverse privacy sensitivity levels have been considered in the FL setup. The HPFL paradigm is evaluated on two representative classification problems: semantic segmentation and emotion recognition. Extensive experiments demonstrate that the HPFL paradigm can: (i) outperform the popular FedAvg by 18.20% in foreground accuracy (semantic segmentation) and 4.20% in F1-score (emotion recognition) under non-i.i.d. sample distributions, (ii) also outperform the state-of-the-art advanced FL algorithms by 12.40%–17.70% in foreground accuracy and 2.54%–4.10%in F1-score, (iii) achieve FedAvg's maximum foreground accuracy 24 rounds sooner, and (iv) incur no extra client-side computation overhead and negligible communication overhead of 5.95% (semantic segmentation) and 0.15% (emotion recognition).

## 3.4 Boosting few-shot learning by self-calibration in feature space

Few-shot learning aims at adapting models to a novel task with extremely few labeled samples. Fine-tuning the models pre-trained on a base dataset has been recently demonstrated to be an effective approach. However, a dilemma emerges as whether to modify the parameters of the feature extractor. This is because tuning a vast number of parameters based on only a handful of samples tends to induce overfitting, while fixing the parameters leads to inherent bias in the extracted features since the novel classes are unseen for the pre-trained feature extractor. To alleviate this issue, we novelly reformulate fine-tuning as calibrating the biased features of novel samples conditioned on a fixed feature extractor through an auxiliary network. Technically, a self-calibration framework is proposed to construct improved image-level features by progressively performing local alignment based on a self-supervised Transformer. Extensive experiments demonstrate that the proposed method vastly outperforms the state-of-the-arts methods.

## 4 CONCLUSIONS

We are very excited to organise the M4MM workshop to foster research and discussion in transversal methodologies and tools for multimedia research. We expect this forum to attract people interested in the foundations of our community, and to open up interesting discussions on good and bad practices when designing tools and methods targeting multimedia applications and related topics. The complete M4MM'22 workshop proceedings are available at: https://dl.acm.org/doi/proceedings/10.1145/3552487 (the link will become active when the proceedings are published in the ACM Digital Library).