APRIL | 2024
Vol. 133 No. 2

MI
MOTION IMAGING JOURNAL

INSIDE:

- AI Image Analysis in Era of Short-Time Viewing
- Automatic Speech Recognition with Machine Learning: Techniques and Evaluation of Current Tools
- AI-Powered Editorial Systems and Organizational Changes
- And More....

AI and the Future of Media Production

Beyond the Frontier of
New Ideas and Applications

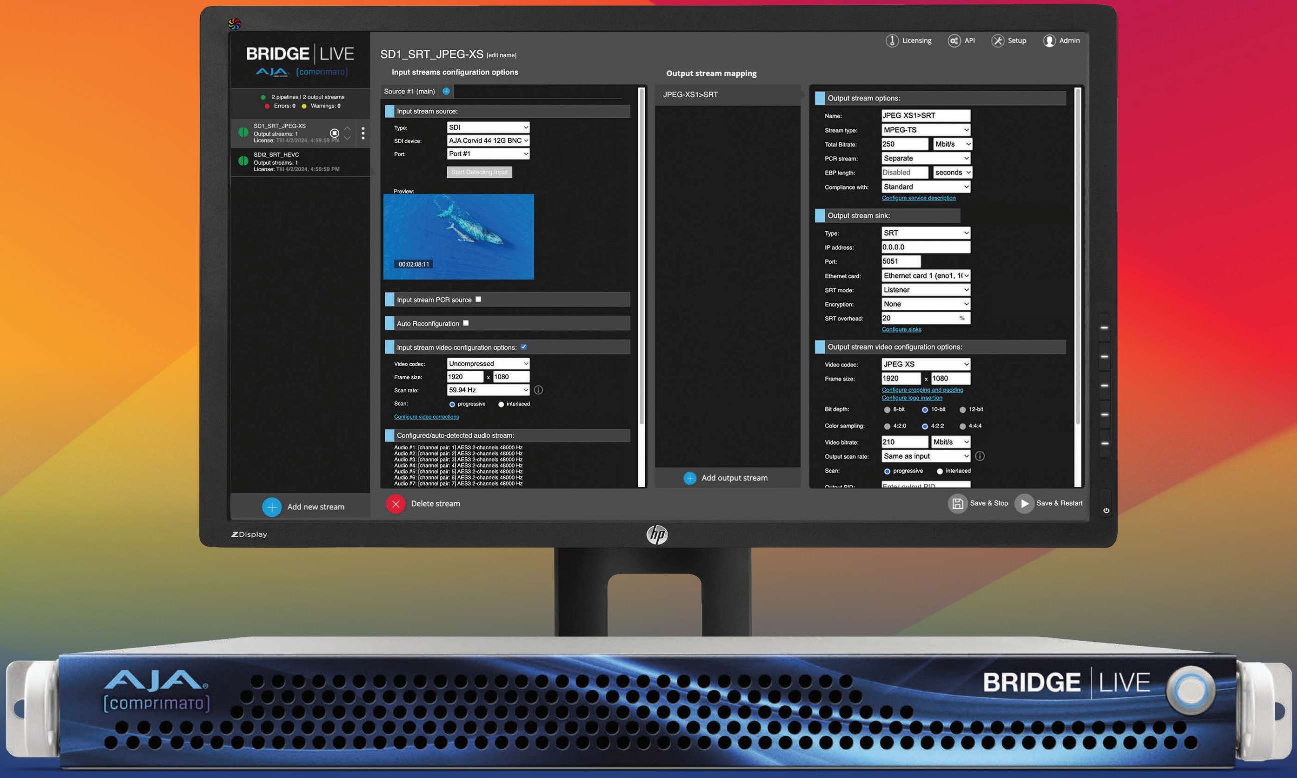
COVERING EMERGING TECHNOLOGIES FOR THE GLOBAL MEDIA COMMUNITY

SMPTE

AJA VIDEO SYSTEMS

BRIDGE | LIVE

The IP Video Workflow Bridge: Now with JPEG XS I/O and RTMP Input



SD1_SRT_JPEG-XS [edit name]

2 pipelines I 2 output streams
Errors: 0 Warnings: 0

SD1_SRT_JPEG-XS
Output streams: 1
License: Till 4/2/2024, 4:59:59 PM

SDI2_SRT_HEVC
Output streams: 1
License: Till 4/2/2024, 4:59:59 PM

Input streams configuration options

Source #1 (main)

Input stream source:

Type: SDI
SDI device: AJA Corvid 44 12G BNC
Port: Port #1

Preview:
00:02:08:11

Input stream PCR source □

Auto Reconfiguration □

Input stream video configuration options: ☑

Video codec: Uncompressed
Frame size: 1920 x 1080
Scan rate: 59.94 Hz
Scan: ● progressive ○ interlaced
Configure video corrections

Configured/auto-detected audio stream:

Audio #1: [channel pair: 1] AES3 2-channels 48000 Hz
Audio #2: [channel pair: 2] AES3 2-channels 48000 Hz
Audio #3: [channel pair: 3] AES3 2-channels 48000 Hz
Audio #4: [channel pair: 4] AES3 2-channels 48000 Hz
Audio #5: [channel pair: 5] AES3 2-channels 48000 Hz
Audio #6: [channel pair: 6] AES3 2-channels 48000 Hz
Audio #7: [channel pair: 7] AES3 2-channels 48000 Hz

Output stream mapping

JPEG-XS1>SRT

Output stream options:

Name: JPEG XS1>SRT
Stream type: MPEG-TS
Total Bitrate: 250 Mbit/s
PCR stream: Separate
EBP length: Disabled seconds
Compliance with: Standard
Configure service description

Output stream sink:

Type: SRT
IP address: 0.0.0.0
Port: 5051
Ethernet card: Ethernet card 1 (eno1, 1(
SRT mode: Listener
Encryption: None
SRT overhead: 20 %
Configure sinks

Output stream video configuration options:

Video codec: JPEG XS
Frame size: 1920 x 1080
Configure cropping and padding
Configure logo insertion
Bit depth: ○ 8-bit ● 10-bit ○ 12-bit
Color sampling: ○ 4:2:0 ● 4:2:2 ○ 4:4:4
Video bitrate: 210 Mbit/s
Output scan rate: Same as input
Scan: ● progressive ○ interlaced
Output PID: Enter output PID

New JPEG XS Encoding, Decoding, and Transcoding

BRIDGE LIVE is the high-end, robust, turnkey system for broadcast-quality, multi-channel, low-latency UltraHD and HD Remote Production and Collaboration, Contribution, Streaming, and Delivery.

The latest software update introduces JPEG XS encode and decode as a licenseable option, plus RTMP as a source input. BRIDGE Live also supports powerful NDI abilities, including encoding and decoding to/from SDI, and realtime H.264/H.265/MPEG-2/NDI file-to-file transcoding.

BRIDGE LIVE is the most flexible and easy to deploy enterprise level IP video bridge for the most demanding workflows.

BRIDGE Live offers seamless compatibility with top streaming CDN services and more.

LIVE YouTube LIVE aws Akamai twitch WOWZA media systems

www.aja.com/bridge-live

Awards Nominations

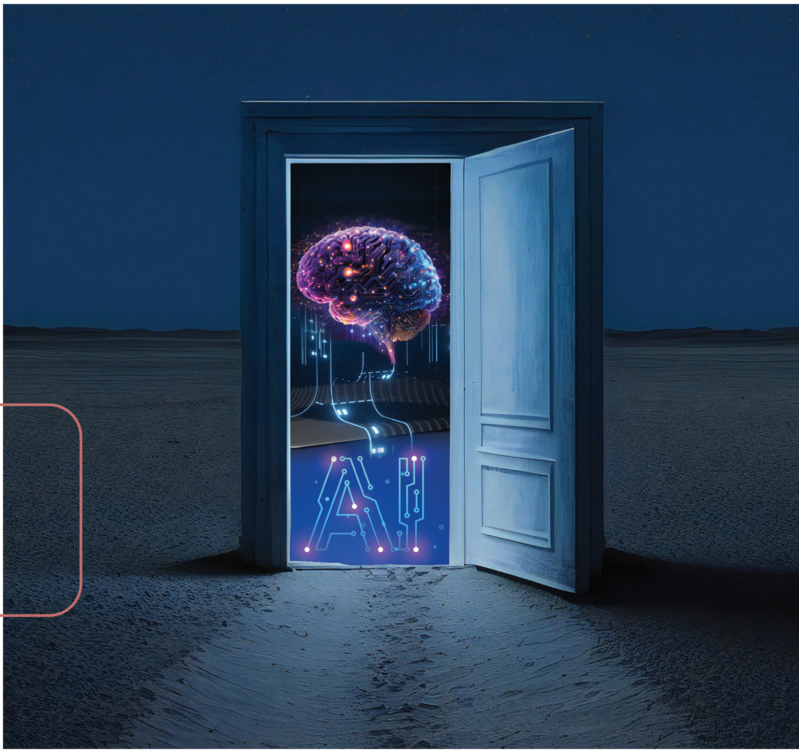
DId you know that any SMPTE member can submit an award nomination?

View our current awards and past recipients on our website, plus download and submit a nomination form. Deadline for submissions is 15 April 2024.

EVOLVE with us
SMPTE

MI
MOTION IMAGING JOURNAL

VOLUME 133 | NUMBER 2 | APRIL 2024



INTRODUCTION

TECHNICAL PAPERS

TUTORIAL

APPLICATIONS/PRACTICES

OFF-TOPIC PAPERS:

1545-0279/24©2024SMPTE

DEPARTMENTS

The Society of Motion Picture and Television Engineers, Inc., 445 Hamilton Ave., Suite 601, White Plains, NY 10601.

EXECUTIVE
Mike Zink
Education Vice President

Jaclyn Pytlarz
Chair/Board of Editors

John Belton
Thomas Edwards
James E. O'Neal
Glen Pensinger
Associate Editors

David Grindle
Executive Director

Steven Bilow
Guest Editor

JOURNAL STAFF
Maja Davidovic
Director of Education
education@smpte.org

Dianne Purrier
Managing Editor, Digital and Print Publications
journal@smpte.org

Anthony Catalano
Executive Account Manager, Advertising Sales & Sponsorships
sales@smpte.org

Jeffrey Victor
Advertising Sales North America
jvictor@smpte.org

SMPTE Motion Imaging Journal (ISSN: 1545-0279) published 10 times per year by the Society of Motion Picture and Television Engineers, Inc., 445 Hamilton Ave., Suite 601, White Plains, NY 10601. Periodi-cals postage paid at WhitePlains, NY, and at additional mailing office. Copyright © 2024 by the Society of Motion Picture and Television Engineers, Inc. No part of this issue may be reproduced without written permission of the publisher. Statements and opinions expressed in articles or editorials in the SMPTE Motion Imaging Journal are expressions of contributors and do not necessarily represent the policies or opinions of the Society. A complete list of SMPTE membership types and benefits is available at www.smpte.org. Advertisements appearing in this publication are the sole responsibility of the advertiser. Subscription rates: SMPTE Members, $35 USD, included in membership dues; print subscription: $1,000 (shipping included); SMPTE owns the copyrights to all published material, unless otherwise stated. *SMPTE Motion Imaging Journal* is registered with Canada Post: CPM# 41003019.

POSTMASTER: Send change of address to *SMPTE Motion Imaging Journal*, 445 Hamilton Ave., Suite 601, White Plains, NY 10601. All rights reserved.
PRINTED IN THE USA

DOI: 10.5594/JMI.2024.IBVN3062
Date of publication: 1 April 2024

SMPTE OFFICERS

OFFICERS

PRESIDENT, 2023–24
Renard T. Jenkins

PAST PRESIDENT, 2023–24
Hans Hoffmann
EBU

EXECUTIVE VICE PRESIDENT, 2023–24
Richard Welsh
Deluxe Media Inc.

STANDARDS VICE PRESIDENT, 2024–25
Sally Hattori
Walt Disney Studios

EDUCATION VICE PRESIDENT, 2023–24
Michael Zink
Warner Bros. Discovery

MEMBERSHIP VICE PRESIDENT, 2024–25
Rosemarie Lockwood
Belden

FINANCE VICE PRESIDENT, 2024–25
Paul Stechly
Applied Electronics

SECRETARY/TREASURER, 2023–24
Lisa Hobbs
MediaKind

EXECUTIVE DIRECTOR
David Grindle
SMPTE Headquarters

DIRECTORS

STANDARDS
Dean Bullock
Florian Schleich
Pierre-Anthony Lemieux
Raymond Yeung
Steve Llamb
Thomas Kernen

EDUCATION
Iris Wu
Monica Brighton
Polly Hickling

MEMBERSHIP
Chris Lapp
John Shike
Zandra Clarke

GOVERNORS

GOVERNORS-AT-LARGE
Chaitanya Chinchlikar
David Long
Jim Helman
John Ferder
Richard Friedel
Stan Moote
Troy English

ASIA/PACIFIC REGION
Michael Day, 2024–2025
Telstra Corporation

Tony Ngai, 2023–24
Society of Motion Imaging Ltd.

CANADIAN REGION
Tony Meerakker, 2024–2025
Meer Tech System

François Bourdua, 2023–24
VS-Tek, Strategies & Technologies

USA - CENTRAL REGION
William T. Hayes, 2023–24
Iowa Public Television

USA - EASTERN REGION
David Long, 2024–25
MAGICSpell Studios, RIT

Dover Mundt, 2023–24
Riedel Communications

USA - HOLLYWOOD REGION
Eric Gsell, 2023–24
Dolby Laboratories

Marty Meyer, 2023–24
Gravity Media

Kylee Peña, 2024–25
Adobe

EMEA, CENTRAL & SOUTH AMERICA REGION
Fernando Bittencourt, 2023–24
FB Consultant

Dagmar Driesnack, 2024–25
Rohde & Schwarz GmbH & Co. KG

USA - NEW YORK REGION
Thomas Mauro, 2023–24
Media Consultant

USA - SOUTHERN REGION
Frank Torbert, 2024–25
WKMG-TV

U.K. REGION
Chris John, 2023–24
Sky UK

USA - WESTERN REGION
Jeffrey F. Way, 2023–24
Quantum

Cassidy Lee Phillips, 2024–25
swXtch.io.

SECTION CHAIRS

ATLANTA
Jack Linder
Linder Consulting

AUSTRALIA
Paul Whybrow
Bodyboard Immersive Experiences

CHICAGO
Jim Skupien
Riedel

DETROIT
R. Thomas Bray
University of Michigan

FLORIDA
Shawn Maynard
Florical Systems

GERMANY
No Section Chair

HOLLYWOOD
Marc Zorn
Marvel Studios

HONG KONG
Terence Yiu
Phoenix Satellite Television Co. Ltd.

INDIA
Ujwal Nirgudkar

ITALY
Alfredo Bartelletti
BLT Italia srl

MONTREAL/QUEBEC
David Beaulieu
CBC / Radio-Canada

NASHVILLE
No Section Chair

NEW ENGLAND
Martin P. Feldman
Unique Scientific, Inc.

NEW YORK
Ed DeLauter
Paramount

NORDIC
No Section Chair

OHIO
John R. Owen
Q Communications, LLC

PHILADELPHIA
Mark Mullen
Millersville University

PACIFIC/NORTHWEST
No Section Chair

POLAND
Kamil Rutkowski
Black Photon Sp. z o.o.

ROCKY MOUNTAIN
Kerry Booth
Signiant

RUSSIA
Suspended

SACRAMENTO
Paul Turner
AJA Video

SAN FRANCISCO
Kent Terry

TEXAS
Curtis Kirk,
PlayMate Enterprise LLC

TORONTO
Lawrence St-Onge
LSO Consulting

UNITED KINGDOM
John Ellerton
BT Media & Broadcast

WASHINGTON, DC
Maciej Ochman
Corporation for Public Broadcasting

STUDENT CHAPTERS BY SECTION

ATLANTA
Clark Atlanta University
Atlanta, GA

AUSTRALIA
Sydney Student Section
Sydney, Australia

DALLAS
Baylor University
Waco, TX

CHICAGO
Purdue University
West Lafayette, IN

FLORIDA
Full Sail University
Winterpark, FL

GERMANY
Universities of Applied Sciences Germany (HdM / HSHL / HSRM)

- Hochschule der Medien | Jan Fröhlich
- Hamm-Lippstadt University of Applied Sciences | Stefan Albertz
- RheinMain University of Applied Sciences | Wolfgang Ruppel

HOLLYWOOD
Los Angeles City College
Los Angeles, CA

Media Arts Collaborative Charter School
Albuquerque, NM

Pasadena City College
Pasadena, CA

USC University of Southern California
Los Angeles, CA

University of Nevada, Las Vegas
Las Vegas, NV

HONG KONG
Hong Kong Design Institute (HKDI)
Hong Kong

Hong Kong Institute of Vocational Education (IVE)
New Territories, Hong Kong

INDIA
Whistling Woods International
Mumbai, India

NEW ENGLAND
Fitchburg State University
Fitchburg, MA

Stonehill College
Easton, MA

NEW YORK
Barnard College
New York, NY

Borough of Manhattan Community College
New York, NY

Bronx Community College
Bronx, NY

Kingsborough Community College
Brooklyn, NY

New York City College of Technology
Brooklyn, NY

Rochester Institute of Technology
Rochester, NY

Rowan University
Glassboro, NJ

William Patterson University
Wayne, NJ

PHILADELPHIA
Millersville University
Millersville, PA

SAN FRANCISCO
De Anza College
Cupertino, CA

City College of San Francisco & San Francisco State University
San Francisco, CA

TEXAS
University of North Texas
Denton, TX

UNITED KINGDOM
Birmingham City University
Birmingham, U.K.

Ravensbourne
London, U.K.

Southampton Solent University
Southampton, Hampshire, U.K.

University of Salford
Manchester, U.K.

University of Surrey
Guildford, U.K.

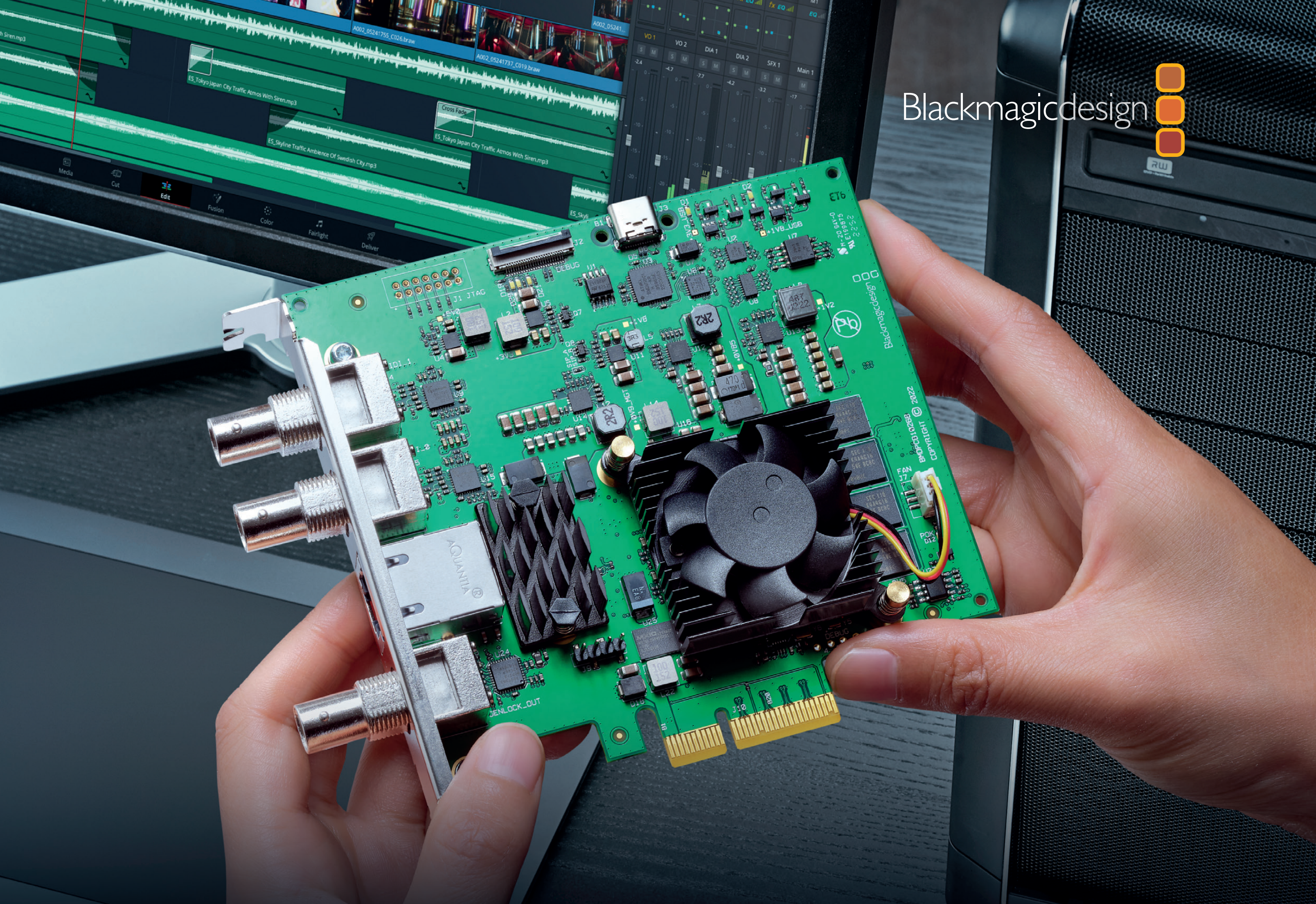
WASHINGTON, DC
George Mason University
Fairfax, VA

Loyola University
Baltimore, MD

Montgomery College
Rockville, MD

BOARD OF EDITORS

Jaclyn Pytlarz, Chair
Ioan Allen
Harvey Arnold
John Belton
Steve Bilow
V. Michael Bove
Wayne Bretl
John Brooks
Dominic Case
Curtis Chan
Greg Coppa
Don Craig
Hamidreza Damghani
Michael A. Dolan
Thomas Edwards
John Ferder
Norbert Gerfelder
Brad Gilmer
Keith Graham
Eric Gsell
Randall Hoffner
Tomlinson Holman
Fred Huffman
Kim Iannone
Scott Kramer
Sara Kudrle
W. J. Kumeling
Michael Liebman
Peter Ludé
Ian MacSpadden
Andrew H. Maltz
Tets Maniwa
Sean McCarthy
Catherine Meininger
Pavel Novikov
James E. O'Neal
Mark O'Thomas
Karl Paulsen
Kylee Peña
Glen Pensinger
Liz Pieri
Charles Poynton
Adam Schadle
Mark Schubin
Gavin Schutz
Thomas A. Scott
Neil Shaw
James Snyder
Steve Storozum
Stephen A. Stough
Simon Thompson
Paul Thurston
J.B Watson
William Y. Zou



The secure way to capture and play video direct to 2110 IP systems!

DeckLink IP cards are the easiest way to capture and play back video directly into 2110 IP based broadcast systems! DeckLink IP cards support multiple video channels plus each channel can capture and play back at the same time. You can build racks of servers generating broadcast graphics, virtual sets, or GPU based AI image processing directly integrated into 2110 IP broadcast infrastructure.

10G Ethernet for Multiple SMPTE 2110 IP Channels

DeckLink IP cards conform to the SMPTE ST2110 standard for IP video, which specifies the transport, synchronization and description of video, audio and ancillary data over managed IP networks for broadcast. DeckLink IP supports SMPTE-2110-20 uncompressed video, SMPTE-2110-21 traffic shaping/timing, SMPTE-2110-30 audio and SMPTE-2110-40 for ancillary data.

High Speed 4 Lane PCIe Connection

With a high speed 4 lane PCI Express interface, DeckLink IP cards feature a PCIe connection to the host computer that's fast enough to handle multiple HD video channels, as well as simultaneous capture and playback on each of the channels. With PCI Express, you get very low latency combined with highly accurate time synchronization to the software that's capturing or playing video.

Supports All SD and HD Formats up to 1080p60!

DeckLink IP connections are multi rate so they support all SD and HD formats. In SD, DeckLink IP supports both 525i59.94 and 625i50 standards. In HD, DeckLink IP cards support all 720p standards up to 720p60, all 1080 interlaced standards up to 1080i60 and all 1080p standards up to 1080p60. There's no complex change over as the cards can switch instantly to a new video standard.

Develop Custom Solutions with the DeckLink SDK

Available free with all DeckLink cards is an advanced developer SDK for Mac, Windows and Linux that you can download free of charge! The DeckLink SDK provides low level control of hardware when you need flexibility, as well as high level interfaces which are a lot easier to use when you need to perform more common tasks.

DeckLink IP HD	$345
DeckLink IP HD Optical	$345
DeckLink IP/SDI HD	$395

Learn more at www.blackmagicdesign.com

CORPORATE MEMBERS

DIAMOND LEVEL

- Apple
- Amazon AWS
- Blackmagic Design, Inc.
- CBS, Inc.
- Deluxe
- Disney/ABC/ESPN
- Dolby Laboratories
- Fox Corporation
- Google
- Paramount Pictures
- Ross Video
- Sony Electronics, Inc.
- Telstra Corporation
- Warner Bros. Discovery

PREMIUM LEVEL

- Academy of Motion Picture Arts & Sciences
- ATOS IT Services UK
- AVID Technology, Inc.
- Bloomberg
- British Telecommunications, PLC
- Imagine Communications
- Microsoft Corp.
- NBC Universal
- Rohde & Schwarz, Inc.

ADVANCED LEVEL

- Absen
- AJA Video Systems Inc.
- AMD
- Belden, Inc.
- Bridge Technologies
- Canon, Inc.
- Dell
- Densitron Technologies
- European Broadcasting Union
- Fuse Technical Group
- Huawei
- Interdigital Communications
- Library of Congress
- Media Solutions)
- Mo-Sys Engineering
- NEP Group
- Novastar
- NTC, A Deloitte business
- Panasonic Corporation
- Qube Cinema
- Red Digital Camera
- Roe Visual Co, Ltd.
- Seagate Technology
- Signiant
- Sky U.K.
- Streamland Media/Picture Shop
- Sudwestrundfunk / ARD
- Texas Instruments
- The Studio - B&H
- Xperi

ESSENTIAL LEVEL

- 4Wall Entertainment
- AOTO
- Appear AS
- Applied Electronics Ltd.
- Arqiva Ltd.
- ARRI, Inc.
- Astrodesign Inc.
- Brompton Technology
- Canare
- Carl Zeiss AG
- CBC Radio Canada
- Chambre des Communes
- Channel 4 Television
- Cisco
- Cooke Optics
- Creamsource
- Dalet Digital Media Systems
- Digital TV Group (DTG)
- Disguise
- Disney Streaming Services
- Diversified
- Ericsson
- Evertz
- EVS/Broadcast Equip
- Extreme Reach
- Fraunhofer
- Grass Valley, Inc.
- ICVR
- IMG Media
- Intel Corporation
- Koninklijke Philips NV
- Leader Electronics Corporation
- Ledyard/Planar
- Matrox Electronic Systems, Ltd.
- Media Links Co., Ltd.
- MediaSilo
- Megapixel VR
- Meinberg-Funkuhren GmbH & Co.
- Motion Picture Solutions
- MLB Advanced Media
- National Association of Theater Owners
- NEC Corporation
- Net Insight
- Nevion
- NHK (Japan Broadcasting Corp.)
- Nvidia
- Pebble Beach Systems
- Perforce Systems
- Phabrix Ltd.
- Pixelogic
- pixit media
- Pixotope
- Portrait Displays
- Quasar Science
- Qube Cinema
- Riedel Communications
- Rosco Laboratories
- Schweizer Radio und Fernsehen
- Sencore, Inc.
- Studio Central
- Synamedia
- Synaptics, Inc.
- Tag VS
- Telestream, Inc.
- Universal Pictures
- V-Nova
- Vu¯
- XR Studios
- Yleisradio Oy
- Zixi

SMALL BUSINESS LEVEL

- 80-six
- Adder Technology
- Adeas, B.V.
- Amphenol RF
- Arista Networks
- ATEME
- Australian Institute of Aboriginal & Torre Strait Islander Studies (aiatsis)
- Aveco
- Barco
- BBC Future Media
- Boland Communications
- BLT Italia srl
- Broadstream Solutions
- Camplex
- Castlabs GmbH
- Chesapeake Systems
- CineCert
- Cobalt Digital
- CST (Comission Superiere Technique de l'image et du son)
- DekTec America
- Deltacast.tv
- Digital Video Group, Inc.
- Disk Archive Corporation Limited
- DSC Laboratories
- Eikon Group Co.
- Eluv.io
- Flanders Scientific
- Fujifilm Inc.
- GDC Technology
- Glassbox Technologies
- IHSE USA LLC
- Imagica Entertainment Media Services, Inc.
- Innovative Production Services
- InSync Technology Ltd.
- Intelligent Wave Inc.
- Internet Initiative Japan
- IntoPIX
- Kino Flo, Inc.
- LAWO
- LG Electronics
- Light Field Lab, Inc.
- Lynx Technik AG
- Macnica Technology
- Marquise Technologies
- Media Tek Inc.
- Merrill Weiss Group, LLC
- Metaglue
- Mole-Richardson Co.
- MTI Film
- Netgear AV
- The Nielsen Company (US), LLC
- NTT Network Innovation Labs
- Original Syndicate
- Panamorph
- Plus 24
- Port 9 Labs
- Qvest Gmbh
- Raysync
- Seiko Epson Corp.
- Showfer Media LLC
- Soliton Systems
- SRI International Sarnoff
- Starfish Technologies
- Strong Technical Services/ Strong MDI
- Sutro Tower, Inc.
- Tajimi Electronics Co., Ltd.
- Tamura Corporation
- Techex Ltd.
- Tedial
- Teledyne Lecroy PSG
- Telos Alliance
- The Family Collective, LLC
- Tokyo Broadcasting System Television, Inc.
- TSL Professional Products Ltd.
- Utah Scientific
- Video Clarity, Inc.
- Visionular

INDUSTRY PARTNERS

- IBC
- ITU-R
- NAB
- NISO

1545-0279/24©2024SMPTE

nabshow.com

2024

Capture YOUR VISION

Register Today!

Use code **MP01** for a **FREE Exhibits Pass!**

Vision. It drives the stories you tell. The audio you produce. The experiences you create. Widen your angle at NAB Show, the preeminent event for the *entire* broadcast, media and entertainment industry. **Discover** a swell of tech and tools across three all-encompassing destinations: CREATE, CONNECT and CAPITALIZE. **Learn** the trends and techniques paramount to your specific role in the content economy. **Network** with innovative people and global visionaries at the forefront of this fast-changing business.

NABSHOW®
Where Content Comes to Life

EXHIBITS: APRIL 14-17, 2024
EDUCATION: APRIL 13-17
LAS VEGAS, NV

Embracing Innovation, Nurturing Talent, and Fostering Diversity

RENARD T. JENKINS

As we enter spring, I'd like to take a moment to reach out to the members of this great Society regarding my presidential platform. At the beginning of 2023, I introduced key components for my presidential platform that would set the stage for the work to be accomplished over the next two years. After ten months, I was able to provide a state of the state for the Society during the SMPTE Media Technology Summit last October. As part of my remark, I acknowledged that while we were still feeling the effects of the strikes in Hollywood, the Society made significant progress regarding member growth and visibility at various educational levels while still focusing on hard science. If you attended the MTS in person, I hope you witnessed the results of the member's hard work.

Now that we are entering the second quarter of 2024 and productions are ramping up again, I would like to review the platform as the components play a role in keeping the Society relevant for years to come.

One of the platform's primary components was ensuring that the Society focused on production sciences, new workflows, technology, and innovation. The wide array and depth of the subjects presented at the 2023 MTS were examples of how we can bring this initiative to the forefront. Another example of this effort is the release of the SMPTE AI report, the first report that thoroughly examines the technology and its potential impact across our industry.

Another platform component was to reinforce SMPTE's participation in tool and standards development for animation, gaming, VFX, and sound. As a result of the member survey, we identified those verticals as growth areas for the organization. To be more engaged with those industry professionals, SMPTE has begun a campaign to engage and better collaborate with the organizations and guilds that represent those groups across the entire media landscape. In the near future, SMPTE will be involved in events that we have not participated in before. I will share more information on those activities with the executive committee and board and solidify those relationships.

We also desire to engage potential members at an earlier age as part of this platform. To move in that direction, we have increased our presence on college campuses and engaged globally with high school and middle school educators. We want to use our influence to engage and support educators as they develop curriculums that will prepare the next generation of engineers, production technologists, and technicians for the industry as it operates today and will in

1545-0279/24©2024SMPTE

the future. This means that we must be ahead of the trends and informed about the direction of our industry. I am happy to announce that we are forming a working group that will have the singular goal of making this component a reality.

WE WANT TO USE OUR INFLUENCE TO ENGAGE AND SUPPORT EDUCATORS AS THEY DEVELOP CURRICULUMS THAT WILL **PREPARE THE NEXT GENERATION OF ENGINEERS, PRODUCTION TECHNOLOGISTS, AND TECHNICIANS** FOR THE INDUSTRY AS IT OPERATES TODAY AND WILL IN THE FUTURE.

Some of you may remember that I stated early on in my presidency that I wanted to ensure that we as a Society serve our members at all levels in their careers. To do this, we need to make sure that we understand the needs of our early-career, mid-career, and late-career members. We have thousands of years of television and film experience across our organization of more than 6,000 members. This allows us to develop programming that improves how our members navigate their careers. We plan to use that wisdom and experience to develop webinars focused on career development and growth for our members. We hope you will participate in this initiative as a contributor and attendee.

This leads me to the next component of my platform: I want to identify and nurture new talent and leaders globally within the Society. We have access to some of the greatest minds in media and entertainment. We also have a global reach that rivals organizations much larger than ours. I want to highlight those members at the local level who are doing incredible work to keep our Society active and visible through engaging events and conferences that bring our community together to discuss and explore transformative technologies that are better changing our industry daily.

In addition, we are actively working on expanding our award slate to include recognition of innovation and up-and-coming talent in the media and entertainment technology career path. Furthermore, SMPTE is connecting with standards organizations in the world of extended reality (XR), virtual reality (VR), and augmented reality (AR) to ensure we remain at the forefront of technological advancements and standardization efforts.

Finally, SMPTE is committed to ensuring inclusivity and diversity in all aspects of our operations to better serve our members, volunteers, and potential members worldwide. This is a key component for an organization's survival. Media production is growing the fastest in Nigeria, while the largest volume of production is in India. This means there are opportunities for SMPTE to impact developing standards, tools, processes, and guidelines in spaces other than just North America. We are a global organization and as president, I want us to seize this opportunity to act and be seen as such. I want to thank every one of you for all you do for the Society, and I look forward to what we accomplish together in 2024.

DOI: 10.5594/JMI.2024.ADUQ8685
Date of publication: 1 April 2024

SPOTLIGHT

SMPTE Member Interview

SMPTE Spotlight:

Joe Addalia

BY RUSSELL POOLE

Media is a vibrant ecosystem that faces constant changes as new technology is introduced. For this ecosystem to survive and thrive, leaders need to integrate this technology in a thoughtful way. That's where Joe Addalia comes in. He is currently the Vice President of Broadcast Technology for Hearst Television Inc., where he is responsible for discovering and implementing new technology. His work involves innovating workflows for News Technology, Broadcast Operations Technology, Media Workflows, and NextGenTV.

"I joined Hearst with the purchase of WKCF-TV in 2006," said Addalia. "The Executive Team's guidance has been tremendous for me throughout the years. Hearst is sincerely a one-of-a-kind company and to be on the inside of the long-range and strategic thinking has helped me develop a view of technology that is not achievable when working short-term or quarterly. This is a perspective which I am truly grateful to have gained."

Addalia joined SMPTE as a Student Member while

CURRENT POSITION:
Vice President, Broadcast Technology for Hearst Television Inc.

PROFESSIONAL ORGANIZATIONS:
SMPTE, Society of Broadcast Engineers (SBE), Advanced Television Systems Committee (ATSC), National Association of Broadcasters (NAB) Television Technology Committee

DEGREES:
Applied Science Degree, Television

"MY HOPE IS TO SEE YOUNG PEOPLE JOIN THE RANKS AND HOPEFULLY BE **AFFORDED THE SAME OPPORTUNITIES** THAT I HAVE HAD FOR CAREER ADVANCEMENTS.

acquiring his Applied Science Degree in Television. He started his career strong, becoming the chief engineer of Newswatch 8 by Adelphia Communications in Ocean County, NJ at the age of 21. “SMPTE standards have always been integral to my technology advancement efforts,” says Addalia. “They have always been about education and standardization, both of which have always played a huge role in my day-to-day.”

Addalia then worked for Press Communications LLC, a Radio and Television Broadcasting Company based in Wall, NJ. He held the position of Corporate Director of Engineering and was responsible for designing and constructing radio facilities in New Jersey and Florida. His biggest project, however, was the creation of WKCF-TV’s studio and transmission facilities, becoming the station’s chief engineer in 1988.

“The leadership at Press Communications showed me that the ‘little guy’ can come out on top in the business,” said Addalia. “I hope I have carried that forward throughout my career with team members close to me.”

Emmis Communications acquired WKCF, and Addalia became the station’s director of engineering while simultaneously acting as the organization’s corporate director of engineering technology. While working there, he implemented Emmis Centralcasting Model for the Emmis TV stations and spearheaded efforts to drive on-air operations through metadata. This was a breakthrough in his career.

During his time at Emmis Communications, Addalia helped research and implement new technologies for 16 TV and 25 radio stations. His work continued when WKCF was acquired by Hearst, adding groundbreaking technologies to systems workflows such as metadata, artificial intelligence, and machine learning. He’s also formed valuable partnerships with Adobe, Bitcentral, Harmonic Inc., Florical Systems, Imagine Communications and TVU. His work earned him a Technology Leadership Award in 2016.

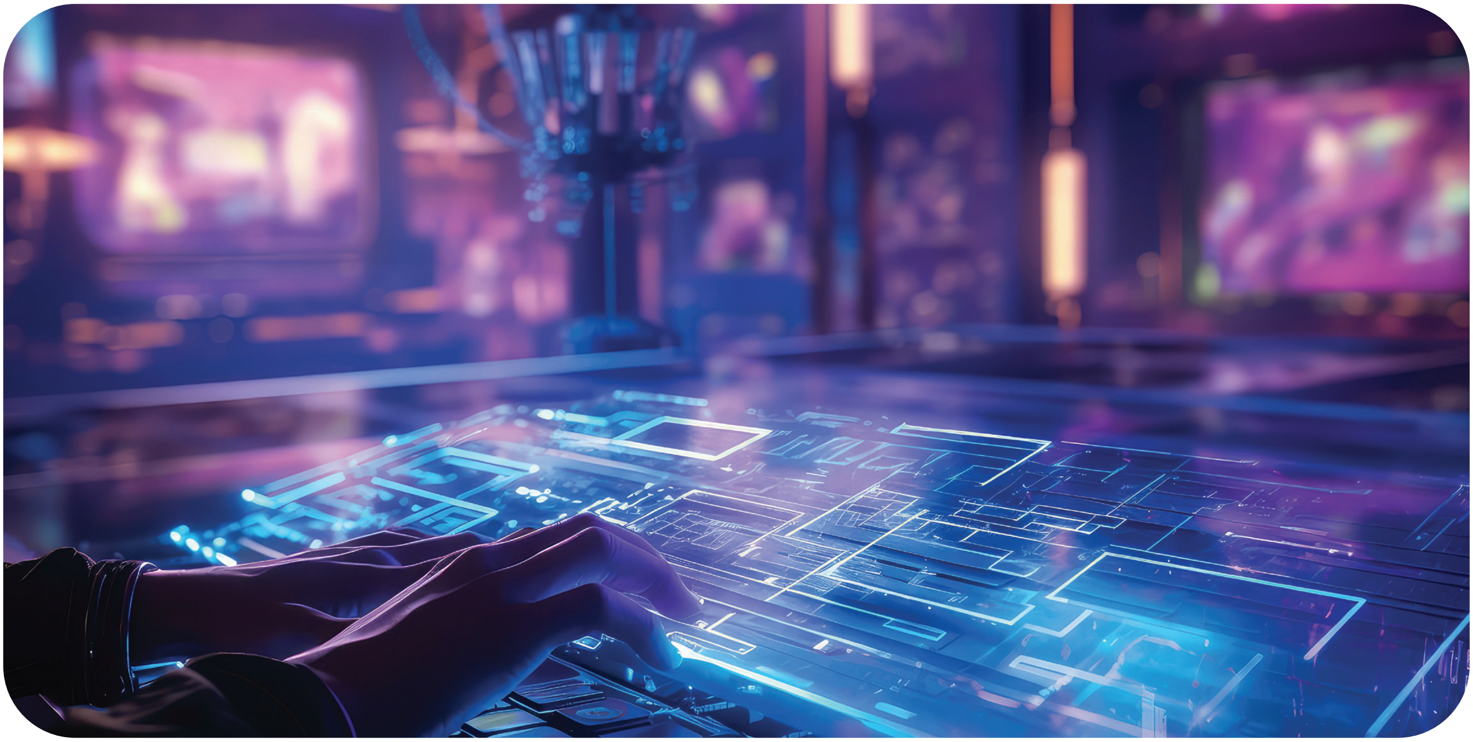
When asked about his lasting impact on the industry, Addalia replied, “The television industry’s ever evolving technology roadmap will extend and iterate for years to come, and it is my hope that my fingerprint on the tech and direction will continue to be the groundwork to chart our company and the industry through the future.”

He also hopes to get students involved as well, stating, “My hope is to see young people join the ranks and hopefully be afforded the same opportunities that I have had for career advancements. Programs like the Marty Faubell Hearst Fellowship and SMPTE Education are key to making this happen.”

New technology is meaningless if it’s not properly integrated. Adding new tech to workflows involves skill, patience, and the determination to see things through. Joe Addalia has all these traits in spades. They are what shaped a successful 40-year career and continue to shape the future of media technology. We are honored to have Addalia as a SMPTE member and manager for the Florida Section.

DOI: 10.5594/JMI.2024.BJIV5618
Date of publication: 1 April 2024

Visit www.joinhearsttelevision.com/fellowships/marty-faubell-fellowship or scan the QR the code for more information about the Marty Faubell Hearst Fellowship program.



NEW AND UPCOMING:

SMPTE Releases Engineering Report on Artificial Intelligence and the Media

SMPTE, in conjunction with the European Broadcasting Union (EBU) and the Entertainment Technology Center (ETC), have released a comprehensive document on Artificial Intelligence (AI) and its effect on the media. The document was the result of a task force on AI standards in media that began in 2020.

"When we started this project in 2020, many saw AI as technically challenging, risky, costly, and even scary," said task force co-chair and AMD Fellow, Fred Walls. "Since then, it has become clearer that AI will transform the media industry from preproduction to distribution and consumption. This report is a great read for those looking to understand AI and how it is being deployed in media, as well as the important roles of standards and ethics."

This report was created to provide media professionals with a background in both AI and Machine Learning (ML). It begins with a technical understanding of the two technologies followed by the effect they will likely have on the media landscape. The report then moves on to examine AI ethics and ends by discussing the role that standards can play in AI/ML's future.

"I believe that AI will continue to see exponential growth and adoption throughout 2024," said SMPTE President Renard T. Jenkins. "Therefore, it is imperative that we examine the overall impact that this technology can have in our industry. That is why the progressive thought leadership presented in this document is so important for us all."

Drafting this report was a joint effort by SMPTE and the ETC with support from the EBU. Members and non-members of SMPTE can access the document for free on the SMPTE website.

Those interested in becoming part of the SMPTE Standards community can find more information at https://www.smpte.org/standards/learn-about-standards-com

DOI: 10.5594/JMI.2024.AUWY9700
Date of publication: 1 April 2024

DOWNLOAD

SMPTE ENGINEERING REPORT
Artificial Intelligence and Media

SMPTE ER 1010:2023

1545-0279/24©2024SMPTE

What's New for Members in 2024!

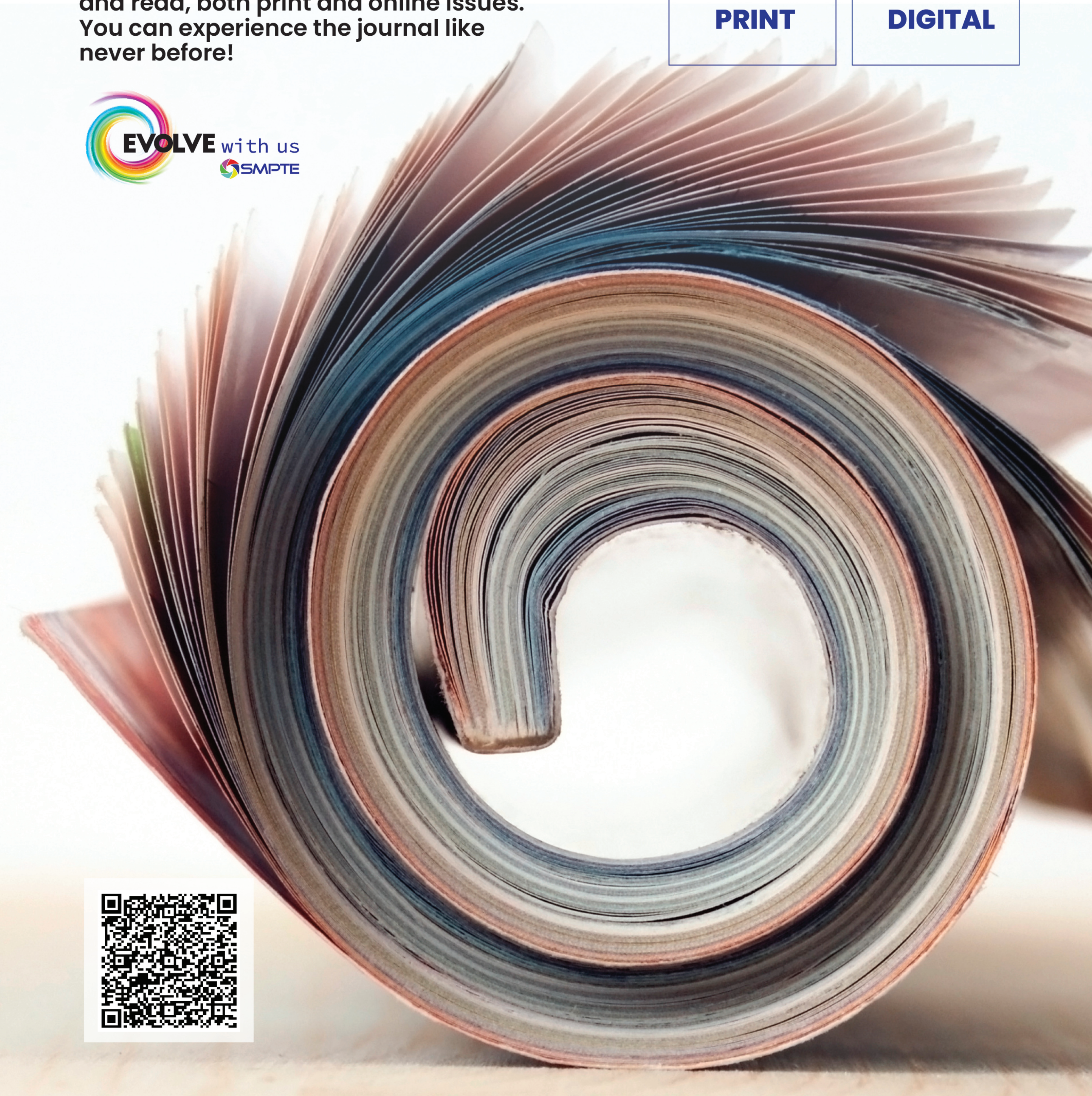
With membership, you will also have access to the new and improved SMPTE *Motion Imaging Journal*, one of the most valuable publications in the media technology industry—Easier to access and read, both print and online issues. You can experience the journal like never before!

EVOLVE with us SMPTE

NEW JOURNAL FORMATS

PRINT

DIGITAL



The Power of Color Symposium: Exploring Color Art, and Technology

BY RUSSELL POOLE

On the morning of 6 February 2024, SMPTE introduced the Power of Color Symposium (POC) to the world. This event showcased the intersection of art and science, new technologies in the realm of color science, and the importance of accurate representation in film and television. Over two days, those at the forefront of color science gave powerful, dynamic presentations that challenged established media experts and inspired future leaders.

While current industry leaders learned from the event, the students of Spelman and Morehouse Colleges, and Clark Atlanta University (CAU), the gracious host of the event, made the symposium a success. These students absorbed all the knowledge offered to them, networked with some of the industry's greatest minds, and thrived when given the opportunity to work on-site at the event.

So, what was presented at the Power of Color? What did the speakers, leaders, and attendees say about it?

Tuesday, 6 February

The event opened with welcoming remarks from SMPTE leadership, members, and our wonderful hosts at CAU,

(L-R): SMPTE President Renard T. Jenkins; CAU students Jacquelynn Dupree and Kennedy Hampton; and Brian Bentley, EdD, CAU Associate Dean of the School of Arts & Science delved into the impact video games have on culture, industry, and education and how the video game industry has greatly impacted contemporary society.

1545-0279/24©2024SMPTE

including the Associate Dean of Arts & Sciences, Brian Bentley, EdD. “The Power of Color Symposium at Clark Atlanta University was Breathtaking,” said Bentley about the event. “This was truly an exceptional opportunity for students, faculty, and industry professionals across the globe.”

Other welcoming remarks were given by Zandra Clarke, transmission specialist at Warner Bros. Discovery; Michele Wright, PhD, SMPTE Director of Business Development and Outreach; Renard T. Jenkins, SMPTE President; Charlene Gilbert, CAU Provost; Andre Dickens, Mayor of Atlanta; and Laquie TN Campbell, Mistress of Ceremonies.

The first session of the day focused on representation and why it mattered. Elfried Samba, CEO and co-founder of the Butterfly 3ffect, spoke on the science and impact of representation, addressing the need for greater diversity in all fields, not just film and television. “Despite enduring an 18-hour flight from Dubai to Atlanta, I approached the event with great optimism and excitement,” said Samba. “It exceeded my expectations. It was amazing to connect with professionals and students who share a genuine enthusiasm for the future of color and its significance in representation and collaboration.”

One such student was Paul Ekomwen, a sophomore and entrepreneur who plans to make a significant mark on the world. “I arrived at this event by chance and got some of the most useful information in my two years in college,” said Ekomwen. “There was so much gold inside of the room and just as much outside in the conversations with speakers. I am thankful to have experienced something that I believe has helped define my path in music, fashion, entrepreneurship, and philanthropy!”

The following session featured Carolyn Calloway-Thomas, PhD, Director of Graduate Studies, African American and African Diaspora Studies at Indiana University, discussing cross-cultural communications as it pertains to viewing skin tones. The session asked, “How do you accurately represent the way you want to be represented?” Calloway-Thomas did not hold back, discussing hard truths regarding representation and requesting everyone listen with open minds and hearts.

Darius Evans, co-president of Georgia Production Partnership, and Saxton Moore, producer and award-winning animation director, then discussed representation in animation. The two leaders discussed how representation could impact a child’s life and create a more vibrant story, which inspired the young crowd intently listening to the session.

The event continued with an interactive session by Christian Epps, founder and CEO of Lights, Camera Diaspora. He demonstrated how different skin tones react to lighting, using equipment that one might find on larger film sets. The discussion challenged people’s traditional opinions on lighting and sought to explain why adequate lighting can represent some accurately and convey the tone a filmmaker is trying to achieve.

The final sessions of day one all dealt with representation in art, including illustration, action, and storytelling. Speakers Keith White of AfroAnimation and Kwame Nyong’o of Savannah College of Art and Design sat with Renard Jenkins to discuss the history of illustration. South Africa native Ingred de Beer, partner and producer of Lucan Animation Studio, and Nigeria-born illustrator and director Shofela Coker discussed their groundbreaking film, *Moremi*. Immix Studios’ award-winning director and producer Lin Tam showed her short, animated film, *Friendship*, and led a fireside chat on how AI will influence animation. It was a powerful way to end the first day.

Session Chair Catherine Meininger of Portrait Displays leads a discussion on “From The Grading Suite: A Colorist’s Perspective” with Grant Reynolds, video colorist at Moonshine Post-Production Studio, and John Petersen, Moonshine Post-Production partner and supervising colorist. They shared their experiences working across a diverse portfolio and discussed current and future evolutions in media color technology.

Wednesday, 7 February

The second day opened with a wrap-up of day one. It led right into the first session with color scientist Catherine Meininger, director of color science at Portrait Displays, who covered everything the SMPTE RIS Color Management working group was working on. This technical session was rife with information, and Meininger made every technical aspect of the presentation easy to follow and understand.

Meininger’s talk flowed into a presentation from Atlanta’s own Moonshine Post-Production. John Peterson, senior colorist partner, and Grant Reynolds, colorist and dailies supervisor, showcased some of their past, current, and future endeavors, and the technology they use to make these projects happen. They also discussed the future of color science and how these technologies will make the job of colorists easier.

Meininger then introduced Ellis Monk, PhD, professor of sociology at Harvard University, who gave a talk on colorism in AI and the Monk Skin Tone Scale. The session detailed how

1545-0279/24©2024SMPTE

Front Row (L-R): Laquie TN Campbell; Michele Wright, PhD; Carolyn Calloway-Thomas, PhD; and Catherine Meininger.

Back Row (L-R): Brian Bentley, EdD; Brian Jenkins; Roy George, PhD; Robert Joseph, PhD; Zandra Clarke; Christian Epps; Ariel Paxton; Grant Reynolds and John Petersen.

AI had a color bias because it mimics humans. To improve AI, we need to improve ourselves. Monk also discussed his Monk Skin Tone Scale and society's bias in skin tones.

After a brief intermission, the event continued with Ariel Paxton, a storyboard artist at Noggin. Paxton discussed the evolving technologies involved in animation and how these new technologies can enhance the storytelling experience. Afterward, Zandra Clark sat down with Robert Joseph, PhD, co-founder of Team MindShift, to discuss how AI can maximize profit and efficiency on a project. Finally, Bentley sat down with Renard T. Jenkins and two of CAU's brightest students to discuss the video game industry.

The last part of the event began with a talk on the creative potential of AI from CAU's professor and chair of the Department of Cyber-Physical Systems, Roy George, PhD, particularly from script development and editing angle. This paired well with the following session on traditional production practices from CAU's TV station manager, Bryan Jenkins.

Bentley and Wright concluded the symposium by acknowledging and thanking everyone who attended, particularly the students. SMPTE and CAU came together a few months ago to form the first SMPTE Student Chapter at an HBCU, resulting in a collaborative force that has changed the lives of many of CAU's finest.

"In the prevailing educational landscape, particularly within the context of Black history, there exists a palpable act of erasure vis-à-vis our stories," said Rileigh Foster, a sophomore at Spelman. "It is imperative for us, as Black creatives, to assume the mantle of responsibility in reclaiming our history and the narratives that have been systematically marginalized. The symposium has served as an enlightening forum, elucidating the expansive possibilities available to us as creatives to redefine the presentation of our historical narrative."

Conclusion

Although the Power of Color symposium has ended, its teachings must continue to make their way through the media tech industry. Film and television serve as ideal avenues to promote education, diversity, inclusion, and growth.

"I am still talking about the SMPTE Power of Color Symposium!" Calloway Thomas, PhD, remarked. "What an enthralling event—so full of fascinating, splendid presentations on a wide variety of topics on the workings of color in the human imagination. The symposium was also a testament to the beautiful people who make SMPTE an influential organization. I enjoyed the symposium in every way and am grateful to have been a part of such a superb event."

The students of CAU, Spelman, and Morehouse have the passion and tools to change this industry, and our responsibility is to help them along the way. The Power of Color Symposium is one step toward a stronger media technology industry, and we are deeply honored that so many students had an opportunity to experience it.

"The Power of Color Symposium was a historic first not only for SMPTE and CAU but for the collective Media and Entertainment Industry as a whole," said Michele Wright, who also served as POC Chair. "The unique array of speakers, participants, and attendees from diverse industries generated a wealth of knowledge, perspectives, and breadth to the importance of accurate depiction across the art, science, and engineering spectrum. This is why it is imperative that, although this symposium is the first of its kind, it must not be the last of its kind. The door to discussing these topics must remain open as our industry continues to embrace inclusivity and purposefully expands its multidimensional reach, impact, and representation."

"I want to thank all of the presenters, participants and especially our hosts, Clark Atlanta University," said SMPTE President Renard T. Jenkinks. "This was a truly informative and enlightening event because of their willingness to share their knowledge. The Power of Color was conceived to be an event where we could bring the hard science of accurate image representation and color science across all content mediums to the forefront of the conversation for all people involved in the creative community...I think the symposium accomplished it's goal."

DOI: 10.5594/JMI.2024.OUFD4983
Date of publication: 1 April 2024

1545-0279/24©2024SMPTE

The Expanding Variety of Media and Entertainment Applications of Artificial Intelligence

BY STEVEN BILOW

The more AI and Machine Learning (ML) experience that our software developers accrue, the more diversely they apply the technology. For several years, this issue of the Journal has experimented with different ways of sharing the breadth of AI-related papers being presented—sometimes, our use cases have been broad and sometimes deep. Last year, we focused narrowly on semantic analysis and covered open-source tools. Before that, our range of examples was more varied. With the wide array of applications where AI is now being employed, presenting diverse coverage seems most beneficial to this audience, and that is what we have done for our 2024 issue.

The past applications we have discussed include AI-assisted captioning, transcoding, storage and retrieval, QA, cybersecurity, image analysis, com-

1545-0279/24©2024SMPTE

THE MORE AI AND MACHINE LEARNING (ML) EXPERIENCE THAT OUR SOFTWARE DEVELOPERS ACCRUE, THE MORE DIVERSELY THEY APPLY THE TECHNOLOGY.

pression, and color matching. As you'll see, our prior coverage represents just the tip of the iceberg. New ideas and applications flow generously.

In the M&E world, we are still just starting to discover the myriad uses for AI and ML. The papers presented here discuss real-world uses for which AI and ML are being employed. What follows is a collection of papers covering Image Analysis, Speech-to-Text techniques, Editorial Systems, querying mixed-format asset storage systems, and Machine Learning approaches to replace hybrid CODECs. We present five AI-related papers and believe they add five new practical applications to the catalog of innovative ideas we have annually been presenting. Here is what you can expect.

In the paper "AI Image Analysis in Era of Short-Time Viewing" by Maezawa, Endo, and Mochizuki, the authors describe an automatic summary video creator that can produce short, social media-appropriate or internet-ready videos from longer form news and other broadcast programs. Their system has been trained on common image composition structures and the camerawork typically used for high-value scenes. The system does not replace highly-skilled creative personnel but it does automatically generate summary videos that have editorial quality remarkably close to that created by creative humans.

In "Using Knowledge Graphs to Enhance Queries Over Heterogeneous Asset Stores," Gonsalves joins Sacilotto and Mathur in a discussion of how knowledge graphs enable the identification of patterns across multiple data sets. They explain how this technology enables data integration and information retrieval between incongruent data stores. The authors focus on the use of graph databases and how their unique data models and query mechanisms empower us to integrate dissimilar data sets irrespective of their original format. The paper also covers ways to use Machine Learning to extract additional information from stored data.

"The Future of Video Compression - Moving Beyond Hybrid Codecs with Machine Learning" by Thomas Guionnet offers an interesting analysis of the benefits and limitations associated with the use of deep learning-based video compression systems. He also investigates practical aspects such as rate control, delay, memory consumption, and power consumption. Following an excellent discussion of the history of Codecs and how ML-based approaches may be used, Guionnet reaches the general conclusion that, even if ML is not quite ready to replace other Codec technology, simple technological progress, and/or some dedicated algorithmic advancements will push solutions forward in that direction.

"Enhancing Content Creation Workflows through Advanced Speech-to-Text Techniques," by Fayan, Montajabi, and Gonsalves surveys and compares several automatic speech recognition models currently used by companies like Google, OpenAI, and Meta. The paper covers voice activity detection, language identification, multilanguage support, and how one characterizes system accuracy. The authors also describe some common challenges, such as inaccuracy resulting from linguistic subtleties. Finally, they describe the role of ASR in media production and dissect the process to offer an insightful analysis of the current state and applications for ASR.

Finally, the paper entitled "AI-Powered Editorial Systems & Organizational Changes" by Arets, Brugman, and de Cooker discusses what they call their "Editorial Portal," an edi-

1545-0279/24©2024SMPTE

torial system whose design considers organizational and technology innovation. They describe design methodologies, like context mapping, to identify relationships that exist between editorial systems and corporate culture. In doing so, they create a system that encourages collaboration and creativity. Journalists from four Dutch regional newsrooms collaborated with the authors so the paper describes not only a new editorial system but a design methodology that directly draws from end-user collaboration.

Beyond these fine papers on our core topic, we include two other stimulating contributions. These days the SMPTE ST 2110 standards are frequently employed in complete facility buildouts. A splendid example of this is the San Ramon, California technology center for the PAC-12 Networks completed in 2023. In the paper presented herein, Neil Kumar and Hieu Ho explain the path followed from operations and workflow analysis through final on-air readiness. Along the way, they examine issues related to Software Defined Network deployment, NMOS integration, troubleshooting, and more. Finally, in "EN17650 – The New Standard for Digital Preservation of Cinematographic Content", Foessel, Sparenberg, Belevantsev, and Lou discuss the requirements and considerations involved in creating the new European standard for long-term cinematic "preservation packages." Their paper covers system architecture, formats, and criteria for selecting requisite metadata. Together, these papers provide stimulating reading whether or not your interest is in this issue's core theme.

AL and ML yearly become more exciting and more relevant to the Media and Entertainment industry and what follows validates that. We hope these papers encourage you to keep considering how the future of AI is far more than those Large Language Models you hear about in the everyday press. Enjoy!

About the Author

Steve Bilow has worked in software engineering, marketing, training, and project management at Tektronix, Grass Valley, Planar Systems, and Telestream. He has been working with video and audio for nearly 40 years. He was first published in the SMPTE Journal in 2001. He is a member of SMPTE's Board of Editors and is a Senior Member of the Association for Computing Machinery where he serves on several USACM committees on artificial intelligence and data privacy. He is based in Portland, Oregon.

DOI: 10.5594/JMI.2024.OHYI1146
Date of publication: 1 April 2024

Design your next products with DekTec

DTA-2172
2x 3G-SDI/ASI
Low profile
Genlock

DTA-2174B
4x 3G-SDI/ASI (1x 12G)
Single or quad-link 4K
Genlock

DTA-2178
8x 3G-SDI/ASI (2x 12G)
8x 12G-SDI with scaling
Genlock

DekTec
www.dektec.com
(303) 318-4298
infousa@dektec.com
Also available:
Satellite, QAM, DVB-T2, ATSC 3.0 receiver and modulator, and ASI I/O

TECHNICAL PAPER

KEYWORDS VIDEO SUMMARIZATION // THUMBNAIL EXTRACTION // NEURAL NETWORK // AI

AI Image Analysis in Era of Short-Time Viewing

By Momoko Maezawa, Rei Endo, and Takahiro Mochizuki

As AI technologies in an era where short-time viewing is preferred, we presented automatic video summarization technologies and a thumbnail extraction technology to distribute summary videos quickly to social media. Many NHK broadcasting stations utilize these systems as tools to activate internet deployments of broadcast content.

1545-0279/24©2024SMPTE

Abstract

In the era when short videos are preferred, broadcasting stations have been enhancing momentums to distribute summary videos of broadcast content on social media. Therefore, we have developed automatic generation systems for news and program summary videos. Using a video summarization artificial intelligence (AI) that has learned the image composition and camerawork typical of important scenes, it is possible to automatically generate summary videos with a high quality close to videos edited by actual program production staff. Our systems include functions enabling users to easily modify the automatically generated summary videos. These systems have been on trial/practical use in many Japan Broadcasting Corp. (NHK) broadcasting stations. The generated summary videos are posted daily on social media. Furthermore, considering a program website is also important content that could boost viewer contact rates, we developed a support system for program website creation using an AI to extract thumbnails automatically. AI-generated thumbnail images can be used to develop program websites wit minimal effort. These technologies can streamline the production of internet content such as summary videos and program websites. Moreover, they will significantly boost internet deployments of various broadcasting programs.

With the increase in consumption of short video clips on the internet, broadcasters are attempting to boost user contact with their content by producing summary videos and distributing them on the internet. Editorial operations to create summary videos require certain specialties and high work costs. Automating the summary video production process for content that is updated daily, such as news programs, is especially desirable. Therefore, we are developing automatic video summarization technologies.

In this paper, we introduce a system that automates the production of news summary videos. Our automatic video summarization technology has the following features. In news footage, the importance of shots is strongly correlated with image composition and camerawork, such as zooming in on important people, panning to show items of interest in detail, and special shooting angles for buildings involved in incidents. We trained an artificial intelligence (AI) system to learn picture-making, such as the image composition and camerawork typical of summary videos produced by skilled editors. Moreover, our technology used the similarity of keywords with an anchor's introduction speech to evaluate the importance of each video segment.

This system summarizes 15–30-min news programs into about 1–2 min with about 20 summary videos produced and distributed daily. Summary videos that used to take a skilled editor more than one hour to produce can now be generated about 10–20 min after the end of the broadcast program, making it possible to produce and distribute summary videos without losing the immediacy of the news.

Regarding common programs not including an anchor's introduction, we developed a practical system to automatically generate summary videos. As with the news video summarization system, the AI on this system learned picture-making typical of important scenes using summary videos created by professional video production staff. This system has also been on trial use in several NHK regional broadcast stations to post summary videos on social media, such as X.

Along with summary videos, attractive thumbnails on program websites can boost the number of accesses to broadcast content. An attractive thumbnail contains an eye-catching image that is representative of the video. We developed a support system for program website creation using AI to select attractive thumbnails from videos. This AI was generated by learning how to compose effective thumbnails that represent the associated video.

The program's website and summary videos can link users to broadcast content, making these technologies a bridge between the internet and broadcasting.

Related Work

Video Summarization

This section describes practical examples of automatic summarization technology in video production sites.

The important scenes to be used in summary videos, such as successful attempts and scores, are clear in sports videos. Therefore, the use summarization technology is being promoted for practical use in various sports events.

1545-0279/24©2024SMPTE

As a practical example for tennis, a summary video generation system developed by IBM was used at the 2018 Wimbledon Championships.[1] Using AI that has learned many point-scoring scenes collected from past game footage, it analyzes the cheers of the audience, movement of the players, scores of the players, etc., and automatically extracts the important scenes in the game. As for golf, at the 2019 Masters Tournament, IBM developed a technology to generate a summary video in a short time.[2] The importance of each shot is determined based on various factors, such as the presence or absence of specific actions (such as fist pumps), facial expressions, changes in the volume and tone of the cheers, and the position of the ball. Each player's importance is assigned automatically, and a summary video of each player is generated automatically. Practical examples for soccer include uniform number recognition, player movement analysis, and automatic summarization technology using techniques such as image trimming in accordance with the results. A summary video distribution service using this technology has started in Italy's Serie A.[3] As for basketball, to popularize and improve the value of the three-player basketball league, "3 x 3.EXE PREMIER," a summary video generation service is in operation that uses technology to detect objects such as scoreboards and recognize displayed characters.[4]

> CONTENT THAT IS **UPDATED FREQUENTLY** IS WELL-SUITED TO THE INTERNET COMMUNITY, WHICH IS SENSITIVE TO TOPIC CHANGES.

In fields other than sports, it is challenging to define important scenes due to differences in editorial staff's points of view and viewers' tastes, and few technologies have been put to practical use.

Hakuhodo DY Media Partners Inc., Tokyo University of Science, and M Data Co., Ltd., have developed a trial version of a system that uses AI to automatically generate summary videos of dramas. They tested it for serial dramas broadcasted from July 2019.[5] However, this approach requires manually assigning metadata to each scene of a TV drama video, based on the performers' utterances and telop contents. Such data is attached to only some of the program videos owned by broadcasting stations. Therefore, we developed a practical technique that can automatically generate a summary video only from a program video.

Thumbnail Extraction

Various techniques have been proposed to select representative images from videos. These techniques use basic image information, such as color histograms, color layouts, texture features related to luminance co-occurrence, motion vectors between frames, and face sizes.[6,7,8] However, the various factors to be considered in selecting images are not always accurately reflected. In addition, a method using a convolutional neural network (CNN), which is the mainstream for tasks such as image classification, has also been proposed.[9] However, this technology is targeted toward YouTube, and because the number of video playbacks is used as training data for CNN, it is difficult to apply it to broadcast video.

The Aesthetic Visual Analysis (AVA) database[10] is a dataset in which images are divided into two classes, high quality, and low quality, based on aesthetic visual analysis by photographers. Jin et al. proposed a neural network (NN) that classifies images into high/low quality using the AVA database as training data.[11] First, the feature extraction network computes the image features, and then the classification network computes the two-class probability distribution from the image features. The selection of thumbnails by program production staff takes into consideration elements related to aesthetic preferences, such as image color and composition. For this reason, we apply this method in our program thumbnail extraction technology. Our technology can improve the efficiency of thumbnail selection work and drive the internet deployments of broadcast content.

News Video Summarization

Content that is updated frequently is well-suited to the internet community, which is sensitive to topic changes. At NHK, expectations are rising for the internet delivery of summarized news to increase the opportunities for viewers to experience broadcast contents. Therefore, we developed an automatic summarization system for news broadcast videos.[12] In this section, we describe the automatic summarization technology for news videos and the functions of the system.

Automatic Split into News Subjects

A typical news program video consists of a plurality of news items. Each news item consists of a part (lead video) in which the studio announcer reads the news, followed by a main story video. Since the content of each news item is completely different, to generate a summary video of the entire news program, it is necessary to automatically divide the news program video into news item units and summarize the main story video of each news item individually.

This section describes the flow of automatic segmentation into news items. First, a news program video is automatically divided into shots, and frame images are sampled from each shot. Then, each frame image is input to the "lead image determination AI" created in advance, and the prob-

1545-0279/24©2024SMPTE

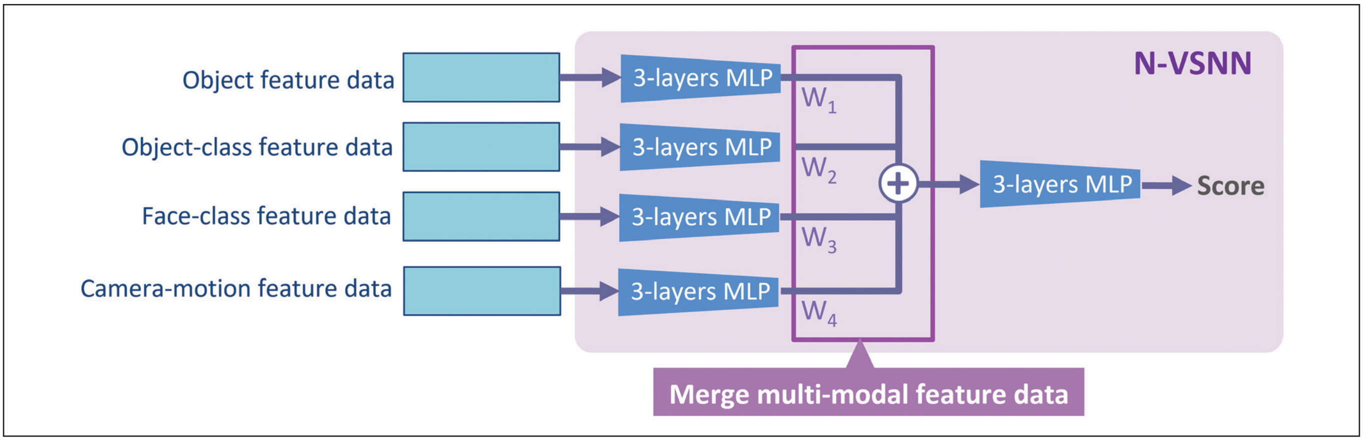


FIGURE 1. News video summarization neural network.

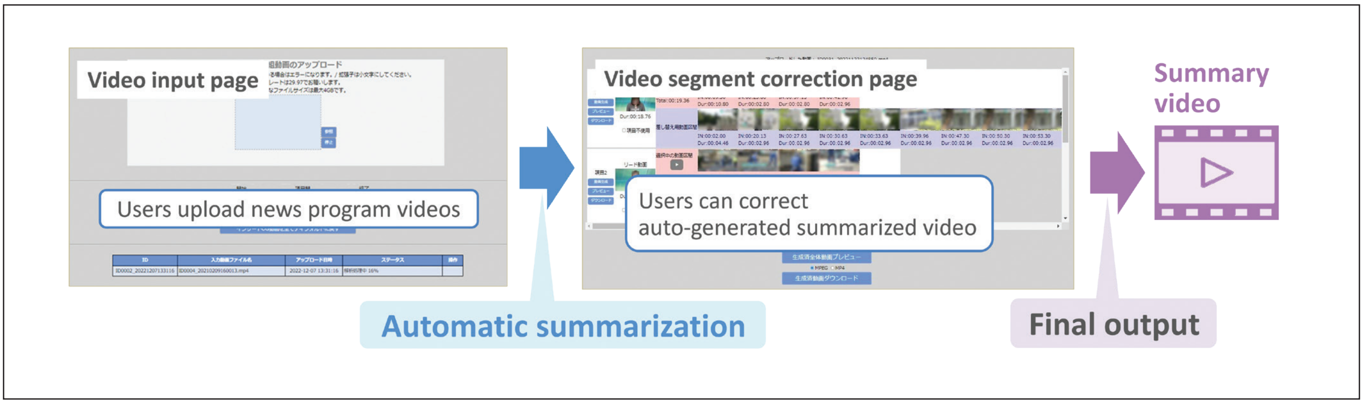

Video input page
Users upload news program videos
Video segment correction page
Users can correct auto-generated summarized video
Summary video
Automatic summarization
Final output

FIGURE 2. Overview of news video summarization system.

ability (score) of each image being included in the lead video is calculated. This process uses a support vector machine that has learned the image features of lead images in various news programs. Next, in each shot, the average score of the images belonging to it is determined. Finally, shots with average scores equal to or higher than a threshold are defined as lead videos. Shot sequences between the lead videos are defined as main story videos and intervals of each news item are determined.

News Video Summarization NN

A video summary of each news item is generated by extracting important video segments from the main story video. We describe the News Video Summarization Neural Network (N-VSNN) for estimating the importance score of each video segment. **Figure 1** shows the structure of N-VSNN. In video production at a broadcasting station, various elements unique to television, such as the type and size of the subject, composition, and camera movement, are considered. Therefore, we made it possible to input multiple modal features to N-VSNN and adopted the following four types of image features (hereafter referred to as "feature datasets") as feature data.

- *Subject feature:* Each image sampled at equal intervals from the video segment is input to the existing trained image classification CNN model,[13] and the feature vector is obtained from the intermediate layer. The average feature vector (2,048 dimensions) for all images is taken as the subject feature.
- *Object class feature:* In the object feature calculation process, the vector (1,000 dimensions) obtained in the same way using the final layer instead of the intermediate layer is used as the object class feature.
- *Face class feature:* From each image sampled in the same way as the subject feature, the face region image detected by Kawai et al.'s method[14] is input to the existing trained face classification model[15] to obtain a feature vector. Then, for each feature vector, a similarity vector is generated using 1,000 face classes created by clustering a large number of face feature vectors. This similarity vector consists of the inner product of the feature vector and the center vector of each class. A face-class feature (1,000 dimensions) is obtained by multiplying and summing the similarity vectors for all face regions by a weighting factor in accordance with the face size.
- *Camera movement feature:* The video segment is divided equally into three sub-segments, the first half, the middle part, and the second half. A vector connecting motion histograms (6 regions x 8 directions) in each sub-segment is taken as a camera motion feature (144 dimensions).

In the previous network, for each modal, 256-dimensional intermediate data is generated through a three-layer multi-layer perceptron (MLP), and the four intermediate data are multiplied by weighting factors W_i (i = 1, ..., 4) and added. When N-VSNN is trained, W_i is updated along with the MLP parameters. Finally, the importance of video segments is output through a post-network consisting of three-layer MLPs.

As training data, we used about 200 summary videos distributed in the past on NHK's news website[16] and news main story videos as sources for each. N-VSNN was generated by training so that a high score would be output for sections of the program video included in the summary video, and a low score would be output for other sections. By using N-VSNN, is it possible to extract video segments with picture-making unique to important scenes in news programs, such as zooming in on important people, panning to show the subject in detail, and special shooting angles for buildings involved in the incident.

Speech Content Score

In the lead video, the announcer gives an overview of the news. Therefore, we consider that the interval where the lead video and the utterance content are similar is important and we introduced the "utterance content score" that expresses the similarity to the summarization process.

In this section, we describe the calculation method of the utterance content score. Keywords are extracted by recognizing the voice of the lead video, and each keyword is vectorized by Word2Vec (hereinafter referred to as a keyword vector). Keyword groups are extracted from each utterance period of the main story video by utterance period detection processing and speech recognition processing, and keyword vector groups are obtained in the same manner as the lead video. For each utterance segment, the degree of similarity between keyword vector groups with the lead video is used as the utterance content score.

News Video Auto-Summarization System

This section describes the News Video Auto-Summarization System developed using the aforementioned processes and N-VSNN. **Figure 2** shows the flow of generating a summary video using this system. The user uploads a news program video to be summarized on the video input page. After uploading, a summary video is automatically generated by the following process.

1. Using the method of Automatic Split into News Subject sections, the news video is divided into news items, and steps 2–7 are performed for each news item.
2. A main story video is divided into shot units, and each shot is divided into fixed-length video segments.
3. For each video segment n, a feature dataset is computed and input to N-VSNN to compute the importance $S_{NN}(n)$.
4. The utterance content score $S_{SP}(m)$ is calculated for each utterance section m using the method in the Speech Content Score section.
5. A total score S(n) for each video segment is calculated using Equation (1).

   $$S(n) = S_{NN}(n) + S_{SP}(m^*)\ R(n, m^*) \qquad (1)$$

   Here, m* is the index of the utterance segment with the highest overlapping rate with the video segment n, and R(n, m*) represents the overlapping rate.
6. The moving image segments are shot out in descending order of S(n). When the total duration of the clipped sections exceeds the length of the lead video, the process ends, and the video segments are connected to generate a summary video.
7. The audio of the summary video is replaced with that of the lead video. Since the content of the lead video is the outline explanation of the news by the announcer, this processing can be expected to improve the explainability of the summary video.

To put the system into practical use, in addition to the function of automatically generating summary videos, it is necessary to perform confirmation and correction work to deploy it to media other than broadcasting. For example, consideration must be given to personal information and the right to use video materials outside of broadcasting. Therefore, we implemented a "video segment correction page" to easily correct the automatically generated summary video. On this screen, it is possible to replace the video segment that constitutes the summary video with another candidate and adjust the IN/OUT points of each video segment. The video summary of each news item that has undergone correction work is connected to a dedicated logo video in between to output the final video summary.

With this system, it is possible to generate summary videos of 15–30-min news programs by automatic processing in about 10–20 min and correction work in about 15 min . Full-scale practical use of this system began in the fall of 2022 and NHK headquarters and many regional broadcasting stations distribute news summary videos generated by this system on social media daily. Furthermore, this system was successfully introduced across all NHK broadcasting stations in the fall of 2023.

Program Video Summarization

For general programs other than news, there is a growing interest in internet development, and there is a demand for a mechanism that automatically generates a summary video. This section describes an NN for program video summarization that we devised[17] and an automatic program video summarization system that we developed using this NN.

Program Video Summarization NN

Figure 3 shows the structure of the Program Video Summarization NN (P-VSNN) for estimating the importance score of each video segment in a program video. Many program videos are long, from tens of minutes to hours. In the process of summarizing long videos, it is common to evaluate the importance of each video segment by considering the video content of the section, nearby sections, and the video content of the entire video. Therefore, P-VSNN is input with feature datasets calculated on multiple timescales, such as nearby shots and the entire video, instead of the feature dataset for the video segment for which the score is to be estimated.

In the network of the first stage, for each modal, feature data of multiple time scales are integrated by a 1-dimensional (1D) CNN and Max pooling, and 256-dimensional interme-

1545-0279/24©2024SMPTE

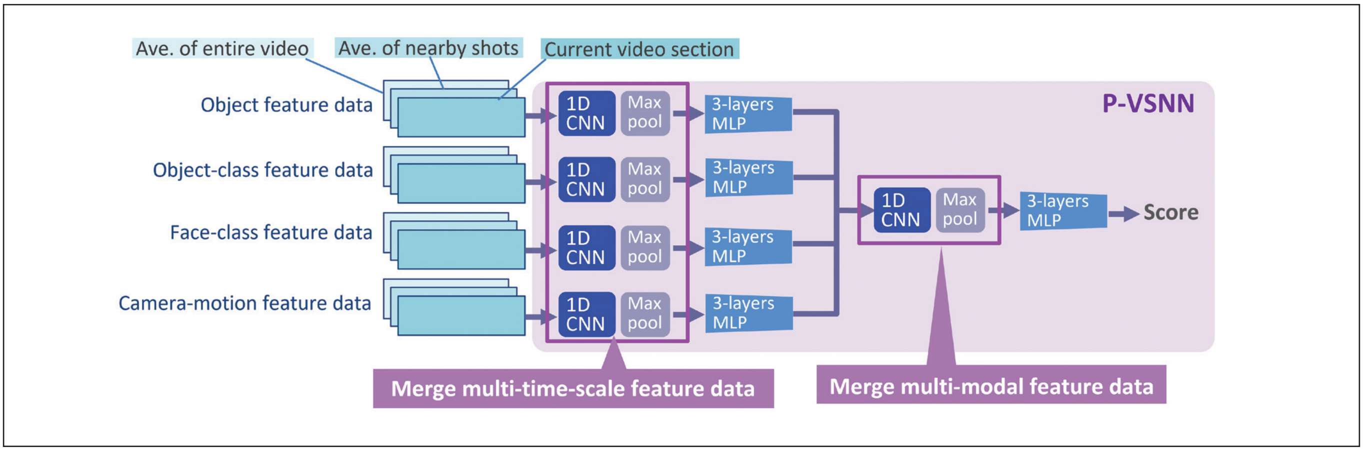


FIGURE 3. Program video summarization neural network.

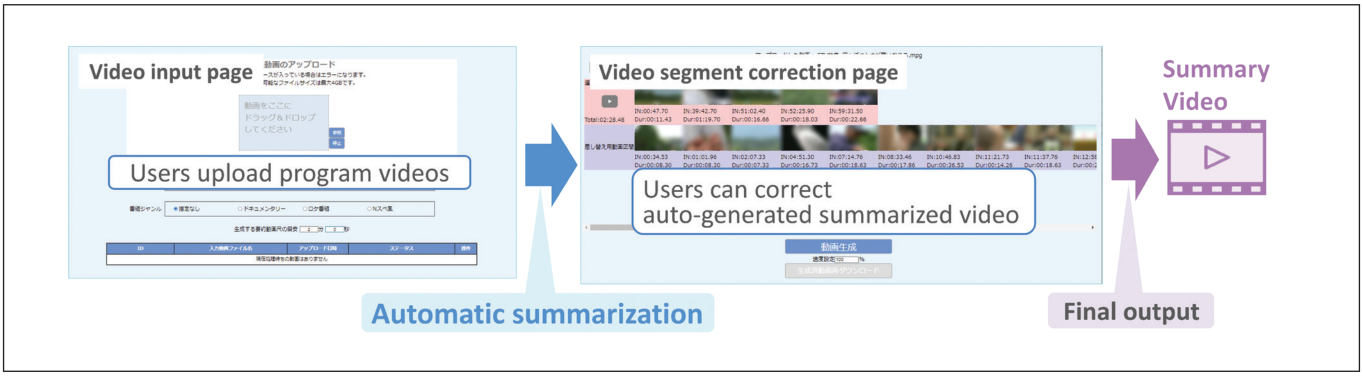


FIGURE 4. Overview of program video summarization system.

diate data are output through a three-layer MLP. In the latter network, the intermediate data of each modal is integrated by a 1D CNN and Max pooling, and the score of the video segment is output through a three-layer MLP.

P-VSNN was trained using dozens of program summary videos created by professional video editing staff, and each program video as learning data. By using P-VSNN, it is possible to extract video segments with picture-making (how to shoot the subject, composition, camera work, etc.) unique to the important scenes of the program video.

Program Video Auto-Summarization System

Using the aforementioned P-VSNN, we have developed an automatic program video summarization system. **Figure 4** shows the flow of program summary video generation by this system. The user uploads a program video to be summarized on the video input page. At that time, it is possible to specify the approximate length of the generated summary video (=T_{SUM}). After uploading the program video, a summary video is automatically generated by the following process.

1. A program moving image is divided into shot units, and each shot is divided into fixed-length moving image sections.
2. For each video segment n, feature datasets at three-time scales (this video segment, average of nearby shots, average of entire video) are computed and input into P-VSNN to calculate importance S(n).
3. The video segments are cut out in descending order of S(n). At that time, to reduce the sense of the incongruity of the sound at the connection point of the moving image section, if the cut-out point is in the middle of the speech section, it can be changed to the start or end point of the speech automatically. An existing library[18] was used to detect speech segments from videos. When the total length of the clipped sections exceeds T_{SUM}, the process ends, and the video segments are connected to generate a summary video.

In this system, we implemented a "video segment correction page" in the same way as the news summarization system, with functions by easy mouse operations such as replacing the video segment that constitutes the summary video with another candidate, adjusting the IN/OUT points of each video segment, and so on. After performing correction work as necessary, the final summary video is output. The user can use this system to generate a summary video with approximately half the program's length and only about 5–10 min of correction work. Currently, several NHK regional broadcasting stations are using this system on a trial basis, and video summaries of cameraman-led programs and other programs generated on this system are being distributed to social media.

Program Thumbnail Extraction

Similar to video summaries, thumbnail images of programs

1545-0279/24©2024SMPTE

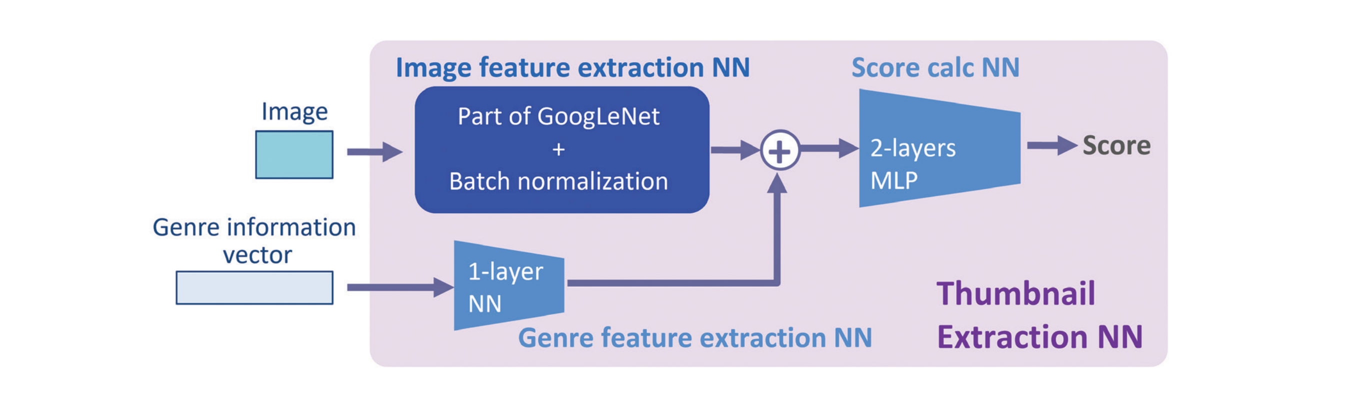


FIGURE 5. Thumbnail extraction neural network.

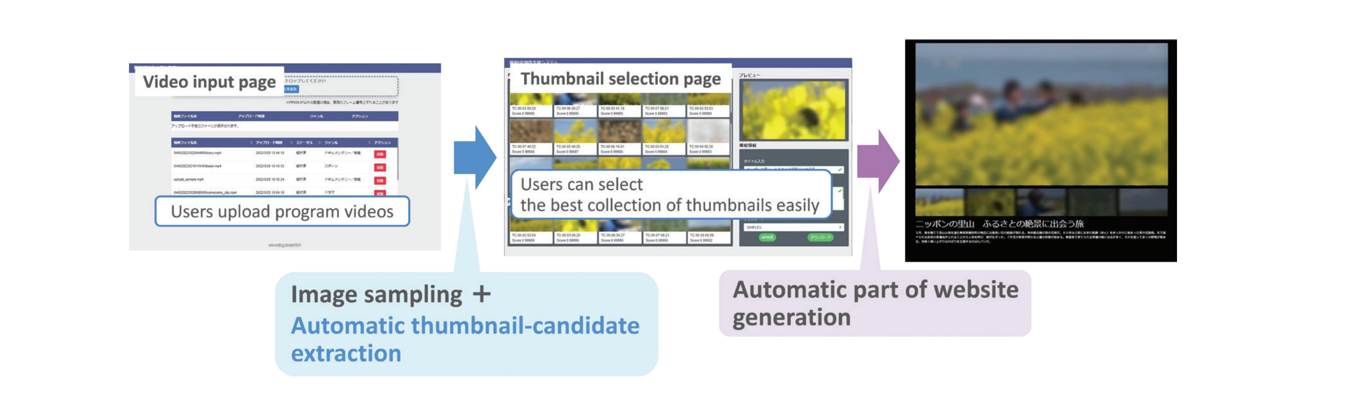

Video input page
Users upload program videos
Thumbnail selection page
Users can select
the best collection of thumbnails easily
ニッポンの里山　ふるさとの絶景に出会う旅
Image sampling +
Automatic thumbnail-candidate
extraction
Automatic part of website
generation

FIGURE 6. Overview of support system for program website creation.

posted on program websites and social media are indispensable items for improving the rate of contact with broadcast content. The presentation of eye-catching thumbnails can be expected to increase the number of viewers accessing programs. However, selecting thumbnails is not an easy task because it is necessary to carefully consider the subject's color, composition, content and so on. NHK disseminates information using many program websites and social media, and improving the efficiency of thumbnail selection work is necessary. This section describes our developed technology for extracting thumbnails and supporting website creation programs.

Thumbnail Extraction NN

The program genre has a significant effect on the thumbnail selection criteria. For example, an image showing the performers is preferred for a drama program, and for a travel program, the beauty of the scenery is given priority. Therefore, we developed a method to select thumbnails from program videos by scoring images using a NN that has learned both images and program genre information.[19] This method makes it possible to select images considering the trend of thumbnails for each program genre.

Figure 5 shows an overview of our developed thumbnail extraction NN. The input is the images sampled from the program video and the genre information vectors representing the program genre (drama, travelogue, variety, etc.), and the output is the score representing the image's suitability as a thumbnail. A genre information vector is a binary vector representing whether a program belongs to each genre by 1/0, and the number of dimensions is 8, which is the number of genres. If the output score equals or exceeds a predetermined threshold value, it is selected as a thumbnail candidate.

Thumbnail extraction NN consists of three networks: an image feature extraction network that computes image features, a genre feature extraction network, and a score computation network.

The image feature extraction network uses the network structure of an existing method[11] that evaluates the visual artistry of images. A structure using GoogLeNet[20] and Batch Normalization[21] generates a 1024-dimensional feature vector. The genre feature extraction network is a one-layer fully connected NN and generates feature vectors (1,024 dimensions) with the genre information vector described later as input. The outputs of the image and genre feature extraction networks are added and input to the score calculation network. The score calculation network, a two-layer MLP, outputs a score between 0.0 and 1.0 using the sigmoid function.

As training data, we used about 6,500 images sampled

1545-0279/24©2024SMPTE

from various program videos and three-level labels (great/good/bad) assigned to each image by program production directors and editorial assistants regarding suitability as program representative images. We set the correct scores for great, good, and bad images to be 1.0, 0.5, and 0.0, respectively, and trained the thumbnail extraction NN using the image and genre information as input. Using this NN makes it possible to extract thumbnails that reflect the image selection skills of professional program production staff.

Support System for Program Website Creation

As previously mentioned, NHK publishes information on many program websites, and there is a demand for a mechanism to streamline the selection of thumbnail images to be posted on the website along with the creation of the website. Therefore, we have developed a system that supports the creation of program websites using the thumbnail extraction NN previously described.

Figure 6 illustrates the process flow of this system. The user uploads a program video, selects the genre of the program, and then a representative image is extracted. Next, shot division of the uploaded video and image sampling processing are automatically performed. The thumbnail extraction NN calculates image scores for the sampled images. Images with high scores are presented as candidate images, and users select the image they want to use. At that time, if necessary, it is possible to input text such as a program outline to be posted on the webpage. Finally, a portion of the website is automatically generated using the selected thumbnail image. Users can preview the generated website and download the HTML files and thumbnail images.

The system will allow for easy creation of a program website and the possibility of further enhancements. In the future, we aim for practical use after trials at the program production site.

Conclusion

As AI technologies in an era where short-time viewing is preferred, we presented automatic video summarization technologies and a thumbnail extraction technology to distribute summary videos quickly to social media. Many NHK broadcasting stations utilize these systems as tools to activate internet deployments of broadcast content. In addition, the program website creation support system using the thumbnail extraction technology is expected to be a practical tool for reducing the cost of creating websites, which is essential to grasp the program easily. In the future, we will work to improve the performance of each technology by retraining AI using data accumulated during trials and practical use and refining the systems based on the needs of program production sites. A mechanism to automatically retrain AI while operating the system could also be necessary.

References

1. How AI picks the highlights from Wimbledon fairly and fast, 2019. [Online]. Available: https://www.ibm.com/blog/how-ai-picks-the-highlights-from-wimbledon-fairly-and-fast/
2. AI-generated Highlights Tell the Story of the Masters, 2019. [Online]. Available: https://blog.video.ibm.com/ai-video-technology/ai-generated-highlights-tell-story-of-the-masters-2019/
3. Serie A agrees AI highlights deal with WSC Sports, 2022. [Online]. Available: https://www.sportbusiness.com/news/serie-a-agrees-ai-highlights-deal-with-wsc-sports/
4. Providing automatic highlight creation service using NTT DoCoMo Co., Ltd., and AI - Easily extract only the highlight scenes you want and reduce the burden of video editing, 2021. [Online]. Available: https://3x3exe.com/premier/news20220601-3/
5. Hakuhodo DY Media Partners Develops an Automatic Drama Digest Video Generation System Together with the Tokyo University of Science: Trial Operation with a New Drama Program Broadcast on Tuesdays by TBS, 2019. [Online]. Available: https://www.inter-bee.com/2019/en/magazine/production/detail/?id=871
6. Y. Gao, T. Ahang, and H. Xiao, "Thematic video thumbnail selection," *Proc. of ICIP*, pp. 4333-4336, 2009.
7. H.-C. Lian, X.-Q. Li, and B. Song, "Automatic video thumbnail selection," *Proc. of Internat. Conf. on Multimedia Technol*., pp. 242-245, 2011.
8. Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes, "To click or not to click: Automatic selection of beautiful thumbnails from videos," *Proc. of ACM International Conference on Information and Knowledge Management*, pp. 659-668, 2016.
9. N. Arthurs, S. Birnbaum, and N. Gruver, "Selecting youtube video thumbnails via convolutional neural networks," *Technical Report*, Stanford, 2017.
10. N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," *Proc. of CVPR*, pp. 2408-2415, 2012.
11. X. Jin, et al., "ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Computer Vision*, 13.2: 206-212, 2019.
12. T. Mochizuki, Y. Kawai, N. Fujimori, M. Maezawa, R. Endo, Y. Asami, "Prototype of support system for news summary video production (in Japanese)," J. *Instit. Image Inform. and Tel. Engin*., 77(2): 262-271, 2023.
13. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *Proc. of ICCV*, pp. 4489-4497, 2015.
14. Y. Kawai, R. Endo, N. Fujimori, T. Mochizuki, "Study of face recognition using deep neural network (in Japanese)," *Proc. of Forum of Information Technology (FIT)*, no. 3, H-003, pp.103-104, 2019.
15. Y. Kawai, R. Endo, N. Fujimori, T. Mochizuki, "Face detection using cascaded convolutional network (in Japanese)," *Proc. of ITE Annual Convention*, no. 3, 22B-1, 2018.
16. NHK NEWS WEB, [Online]. Available: https://www3.nhk.or.jp/news/movie.html
17. T. Mochizuki, Y. Kawai, "Program video summarization by fusion of multi-modal/timescale features using 1D-CNN (in Japanese)," *Proc. of IEICE General Conference*, D-12-26, 2022.
18. Real Python, The Ultimate Guide to Speech Recognition [Online]. Available: https://realpython.com/python-speech-recognition/
19. M. Maezawa, R. Endo, N. Fujimori, T. Mochizuki, "A study on the selection of thumbnail images considering a genre of TV programs (in Japanese)." *IEICE Technical Report*, vol. 121, no. 155, PRMU2021-15, pp. 48-51, 2021.
20. C. Szegedy, et al., "Going deeper with convolutions," *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognotion*. (CVPR), 2015.
21. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proc. 32nd International Conference on Machine Learning (ICML)*, 448-456, 2015.

About the Authors

Momoko Maezawa received BE and ME degrees from Keio University, Japan, in 2017 and 2019 respectively. She is a research engineer with the Science and Technology Research Laboratories, NHK (Japan Broadcasting Corporation).

Rei Endo received BE, ME, and PhD degrees in engineering from Keio University, Japan, in 2008, 2010, and 2013, respectively. He is a research engineer with the Science and Technology Research Laboratories, NHK.

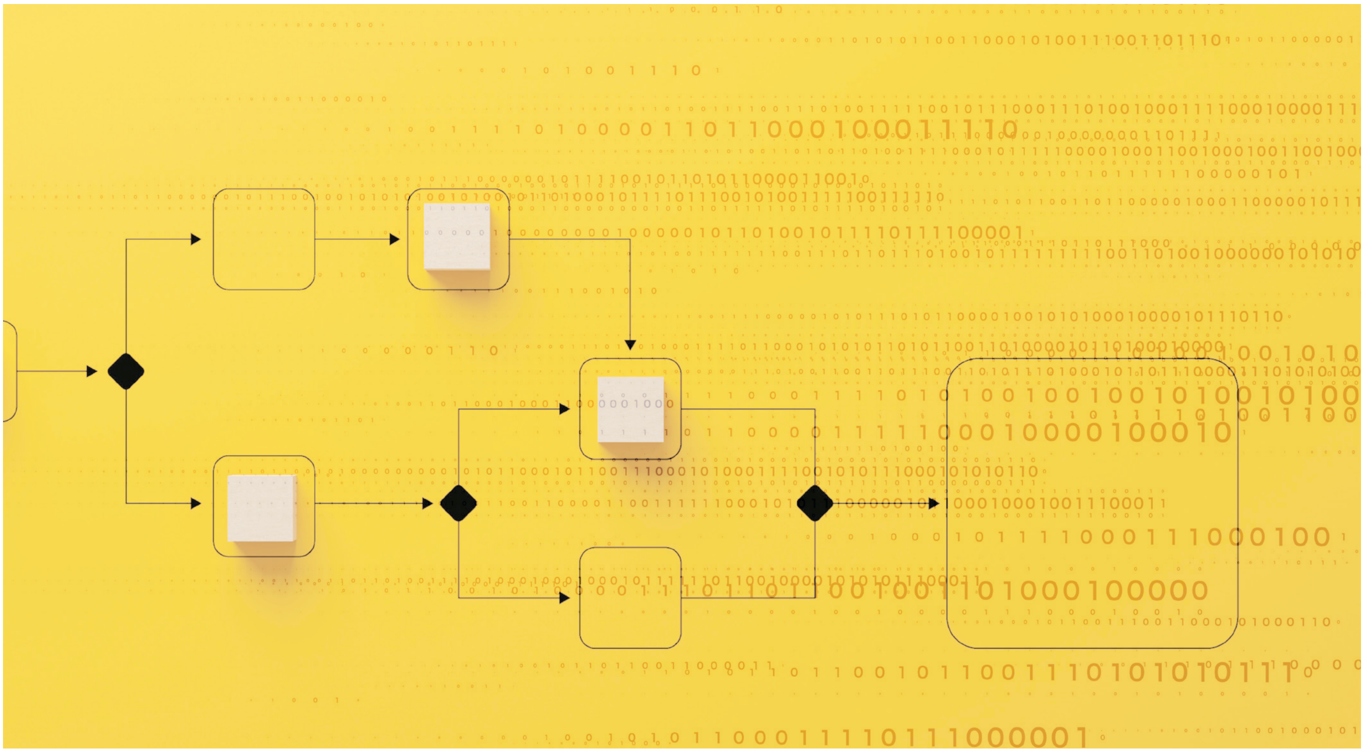
Takahiro Mochizuki received BE, ME, and PhD degrees in engineering from the Tokyo Institute of Technology, Japan, in 1994, 1996, and 2010, respectively. He is a senior research engineer with the Science and Technology Research Laboratories, NHK.

Presented at IBC 2023, Amsterdam, The Netherlands, 15-18 October 2023. This paper is published here by kind permission of IBC. Copyright © IBC.

DOI: 10.5594/JMI.2024/BZSK5411
Date of publication: 1 April 2024

TECHNICAL PAPER

KEYWORDS KNOWLEDGE GRAPHS // DATA INTEGRATION // SEMANTIC UNDERSTANDING // MACHINE LEARNING



Connecting Asset Stores with Graph Databases

By Roger Sacilotto, Shailendra Mathur, and Rob Gonsalves

Knowledge graphs can effectively represent entities and relationships with meaning defined by semantic specification. The structures of these entities and relationships can be processed in a way that provides unique value, especially when entities exist across multiple heterogeneous databases. This aggregate data layer allows workflows to query the graph to find entities based on the relationships and use them to fulfill processing requirements.

1545-0279/24©2024SMPTE

Abstract
As media organizations accumulate diverse data from various sources, effective data integration and retrieval becomes increasingly critical. Graph databases offer a flexible and efficient platform for integration by leveraging graph-based data models and querying mechanisms. Graph databases can seamlessly integrate disparate data sets regardless of their original format or structure. Knowledge can be added to the graph through the application of semantic meaning to the data. Machine learning (ML) algorithms can extract additional information from data to enhance knowledge graphs (KGs). This paper demonstrates how knowledge graphs offer unique capabilities for understanding patterns over collections of large data sets.

Data is stored across multiple databases in various production environments, including those utilizing machine learning (ML), each with unique schemas and rules. This scenario is prevalent in film and television production, broadcasting, and machine learning systems. These environments encompass a wide array of data, from scripts, visual effects (VFX), and media to audio, video, and machine learning-generated text assets. Understanding and documenting the interdependencies among these diverse data elements is essential for creating efficient workflows that integrate different datasets.

It's crucial to identify and process data dependencies in workflows that bridge these datasets. For instance, a film production might need to coordinate a VFX shot, requiring an understanding of how the green-screen footage relates to VFX models. This involves managing data and collaborating across various teams and systems, including those handling outputs of machine learning.

To find connected assets, one usually needs to search for linkage values in databases. Asset retrieval can become challenging if these values aren't indexed or embedded in complex data structures. Graph database technology can streamline this process by creating a linkage database, facilitating easier coordination.

The Semantic Web[1] (Linked Data[2]) concept, used in platforms like DBPedia and Wikidata, employs the Resource Description Framework (RDF)[3] to link web resources. While initially meant for public internet data, this concept also applies to other areas, including environments leveraging machine learning. For example, Movielabs,[4] a research organization, has developed an ontology for film and TV production data. The ontology can be extended to include machine learning-generated data, to unify the structure across different systems.

This report delves into using knowledge graphs to create a data layer that represents and interconnects objects across various data stores, including those driven by machine learning. Graphs enhance workflows and unlock the potential of disconnected complex data, including systems heavily reliant on machine learning.

Data Integration

Modern media workflows usually involve various kinds of data that are incorporated into film, television, or broadcast news workflow. These data elements are combined in projects where a complete view of the data can be useful for content management. Examples of relevant scenarios include:

- Collecting related but independently stored data across systems, in many different forms and types, for archival purposes. Data can include video, audio, graphics, text, closed captions, descriptive metadata, and business information.
- Using reference counts to know when elements are no longer needed.
- Finding the full usage scope of licensed content.
- Identifying overuse of specific content and finding fresh alternatives.

Graph databases can serve as an aggregation layer connecting elements hosted in different "native" databases. The ability of entities to refer to anything with a string-based access key and the arbitrary connectivity that can be encoded provide a system that can conform to a wide variety of data environments.

Example Use Case

To explore the aggregation of heterogeneous data stores in a graph, consider the following hypothetical scenario (the story is true; all names have been changed).

- New England news station WGPH is covering a story about two swimmers caught in a rip current at Shoreville Beach, New Hampshire.
- The story is run in the 6 PM broadcast, anchored by

1545-0279/24©2024SMPTE

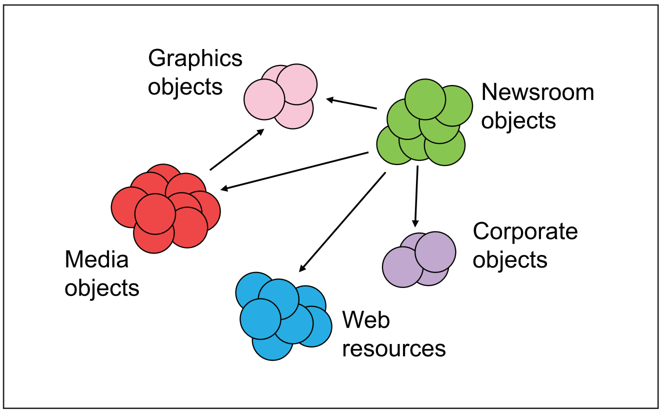


Figure 1. Abstract story data graph.

John Jones and reported by Janet Smith. The story data and 6 PM rundown are stored in the newsroom management system.

- Janet reported from the beach, interviewing Alice Wilson and Will Williams.
- Alice is a bystander who went into the water to help one of the swimmers.
- Will is the assistant fire chief in Shoreville. He commented on the status of the other swimmer and also talked about swim safety.
- A drone and a videographer recorded video. The recorded video is ingested into a content management system.
- The video is edited into a package and added to the story data.
- Lower-third graphics were created in a graphics management system, timed to the package, and added to the story data.
- John introduced the story and tossed it to Janet. Janet set up the package and was back live when it ended.

A graph database was created with elements representing the entities mentioned above and relationships among the elements. The data was not crafted with a carefully designed schema in mind, it only contains enough information to demonstrate the concepts. **Figure 1** shows an abstract depiction of the graph elements and connections to provide a visual concept of the different pieces of data that contribute to this story.

Each circle is an object that represents an entity. The entities are grouped by their type and where they are stored. This scenario, the entities involved in it, and the relations between the entities will be discussed later. First, a graph tutorial should provide some context.

Knowledge Graph Basics

What is a Knowledge Graph?

The term *Knowledge Graph* is defined differently across domains, but generally refers to *data intended to accumulate and convey knowledge of the real world*.[5] Note that the term "real world" is often meant to refer to things outside of the computer domain, but for purposes of this document, it also includes things within the computer domain, e.g., objects that only exist in a database or a file system.

Graphs are represented by graph-structured data models. These models consist of nodes and edges. Nodes represent entities or values. Nodes can be connected to other nodes where a connection is seen as a relationship between two things. The connections are also known as predicates[6] in formal definitions. Connections between entities are represented by edges between nodes. **Figure 2** shows nodes representing three entities:

The images displayed in **Fig. 2** serve only to give an idea of the entities. From left to right, the entities are the Louvre Museum in Paris, the painting *Mona Lisa*, and Leonardo da Vinci. We can connect specific information about the entities, like name and birth date. We can also establish how the entities are related to each other. The information and relationships provide context that helps with understanding. Without context, the text "Mona Lisa" could mean the painting, the song, or a character in a movie. With context, we can know that the entity shown above is positively the painting created by Leonardo da Vinci and displayed in the Louvre. **Figure 3** shows a fuller depiction of the nodes with some edges that provide context:

Edges are labeled to describe the nature of the relationship. In knowledge graph systems, the edges are typically directed from the subject node to the object node. For example, the edge labeled *CreatedBy* originates at the Mona Lisa node and ends at the Leonardo da Vinci node. The data formatted as strings or dates (e.g., "The Louvre") are value nodes also known as literals. These value nodes cannot be the subject of any edge, only the object.

Entity nodes and edge labels must have a unique identity. Identity forms a basis for knowledge. If it is important to understand the location of an object to know exactly what that object is, then the relation label between the object and its location must be known as well. In this case, the *LocatedIn* label

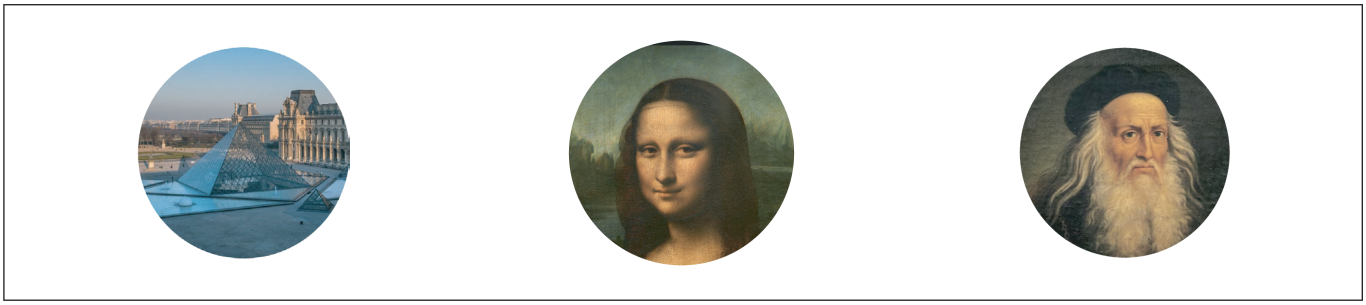


Figure 2. Three entities.

1545-0279/24©2024SMPTE

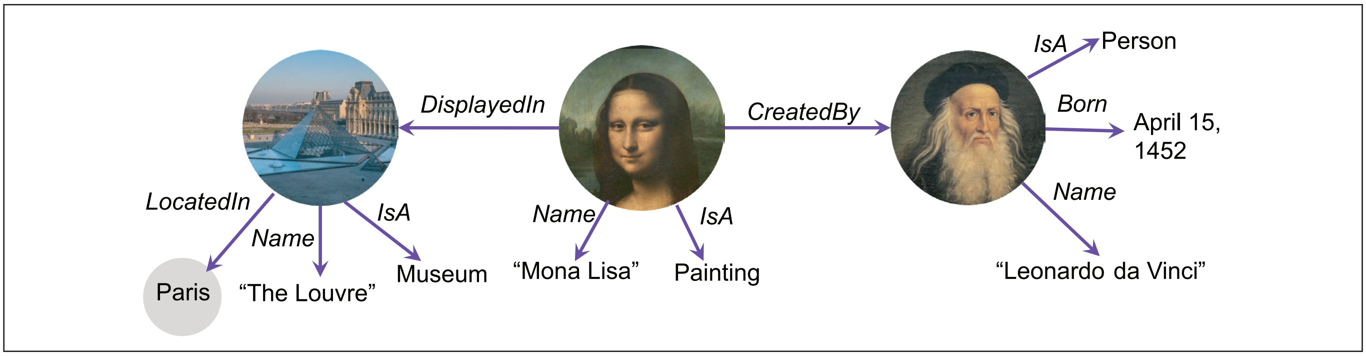


Figure 3. Nodes and edges.

must be precisely defined and cannot be ambiguous, likewise for nodes. The node discussed that represents the Mona Lisa in the painting cannot also represent the Mona Lisa in the song. If a node is ambiguous, then knowledge is diminished.

What does this additional information do to provide knowledge? First, notice the *IsA* relations on the entities. For now, consider the *IsA* relation to connect an entity to an object representing a type of entity. This provides information as to the kind of thing that the entity is. In this case, the da Vinci node is the person, not a painting of the person, and the Mona Lisa is the painting, not the song. The entity type is part of the context that gives meaning to the entity node. For example, improved confidence in person entities can be achieved by considering related contextual elements such as birth dates, birth locations, accomplishments, and family members. This concept applies to all kinds of entities. Increased knowledge about the relevant data can lead to greater confidence.

Identity and Semantics

As mentioned previously, the unique identity and meaning of nodes and edges are critical for knowledge. However, identity management can be complex.

Identity

Knowledge graph identifiers must be encoded into values supported by the graph database software. Strings are a universal datatype and are typically used to represent identity. Given that linked data is a popular method of implementing knowledge graphs, we should start by examining how identity is managed.

As mentioned earlier, RDF is a specification for representing resources on the internet. RDF requires an international object nodes and relations to be identified by an Internationalized Resource Identifier (IRI).[7]

IRIs effectively reduce to a choice between a URL (e.g., http://smpte.org) and a URN (e.g., urn:ietf:rfc:3987). For linked data, URLs are generally preferred because they can be used to fetch data. For example, making an HTTP GET call to the URL http://dbpedia.org/resource/Leonardo_da_Vinci will return RDF data from DBPedia related to Leonardo da Vinci.

A primary benefit of URLs is that they are "self-minting," meaning that an organization with a domain name can manage its namespace of identifiers without registering with a naming authority, e.g., IANA.[8] URNs are less flexible because they require a registered namespace, thus increasing the management burden.

RDF allows the use of prefixes to shorten URLs. For example, the base URL of https://dbpedia.org/resource/ can be assigned to the prefix *dbr*. In some representations (e.g., Turtle,[9] RDF-XML) the resource http://dbpedia.org/resource/Leonardo_da_Vinci can be shortened to dbr:Leonardo_da_Vinci. Prefixes are used in documents containing graph information and are only valid within the document's scope, without the need for global definition. Once read into a database, the identifiers are expanded, and the prefixes are discarded. For example, the following shows a small subset of the data retrieved from the DBPedia HTTP call mentioned earlier in the Turtle format.

```
dbr:Leonardo+da_Vinci a yago:Architect109805475,
    owl:sameAs dbr: L\u00E9onard_de_Vinci .
```

In the data shown, nodes and edges are all represented as URLs. Turtle uses prefixes to shorten URLs. For example, the prefix "dbr" is mapped to "http://dbpedia.org/resource/. The prefix mapping is added to the rest of the identifier to get the full URL. Note that the existence of a URL does not guarantee that the URL can be used to fetch data from that location.

Semantics

Simply put, semantics is the application of meaning to data. Meaning is conveyed with a specific semantic assertion to an entity or edge in the graph. The assertions are specified in an ontology using an ontological language such as RDF-Schema,[10] Simple Knowledge Organization System (SKOS)[11] or the Web Ontology Language (OWL).[12]

Semantic assertions include categories such as classification, taxonomies, and inference. The value of using accepted ontological languages is the consistent interpretation of meaning. Recall the Leonardo da Vinci entity used in examples above. The fact that the entity represents a "Person"was asserted by having an edge labeled *IsA* connecting the da Vinci entity to a node labeled Person. This example was abstract but shows the need to have agreement on the labels. The term "IsA" is not precise, and without a formal definition, the certainty that meaning can be communicated is low. Likewise, the term "Person" is from a literal string point

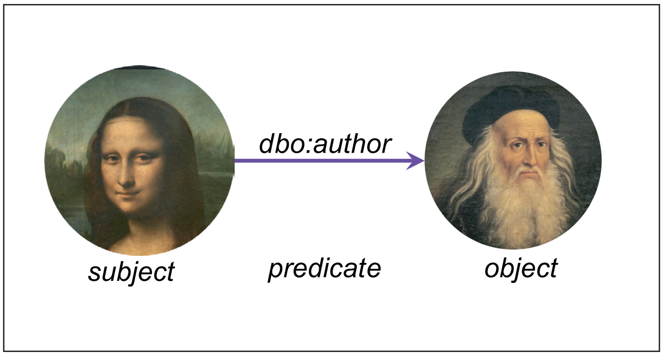


Figure 4. RDF triple.

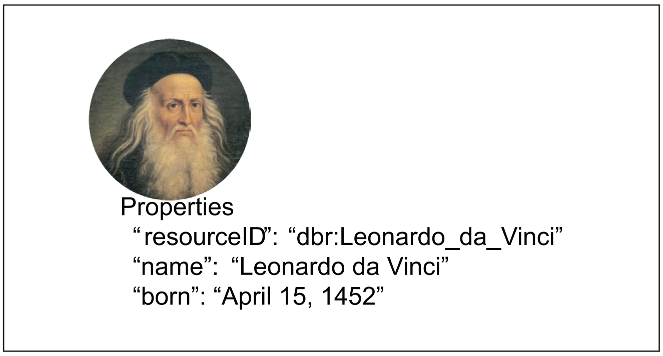

Properties
" resourceID": "dbr:Leonardo_da_Vinci"
"name": "Leonardo da Vinci"
"born": "April 15, 1452"

Figure 5. LPG node.

of view. In RDF and OWL, this is addressed by requiring that entities and predicated be represented with IRIs. The DBPedia record for Leonardo_da_Vinci shows the following triple:

dbr:Leonardo_da_Vinci rdf:type foaf:Person

The predicate rdf:type is the standard way of indicating that the subject entity is a member of a class of things. In this case, the class is identified by Person in the Friend-of-a-Friend (foaf) namespace. Software that knows about these resource identifiers can establish confidence that the Leonardo_da_Vinci resource refers to the Person, not a painting of da Vinci.

Graph Implementation Types

While several graph implementation styles have been created, most have faded away,[13] leaving two dominant styles: Directed Edge-Labelled Graphs, and Labelled Property Graphs.

Directed Edge-Labeled (DEL) Graph

A DEL graph is a set of nodes and a set of directed labeled edges between those nodes. RDF Graphs are a very common form of DEL graphs and will be used to explain their characteristics.

RDF is based on records called triples that contain a source identifier, a predicate and an object as shown in **Fig. 4**.

Here, the triple in DBPedia can be represented as follows:

(dbr:Mona_Lisa, dbo:author, dbr:Leonardo_da_Vinci)

As mentioned above, an object can be a resource identifier or a literal value. For example, using a standard predicate for describing the title of a work, a triple for the name of the Mona Lisa can be represented as:

(dbr:Mona_Lisa, dc:title, "Mona Lisa"@en)

This triple states that the English-language title (or name) of the entity dbr:Mona_Lisa is "Mona Lisa." In Italian, the triple would be the same except for the object value, which would be "Gioconda"@it.

The requirement to store everything (entities and literals) in triples simplifies the internal data structures but makes some concepts a little more complex. A compound literal value, for example, cannot be implemented straightforwardly. Imagine a "coordinates" value with longitude and latitude. The long/lat values must either be encoded in a single string, or an anonymous "blank node" would have to be created to hold the longitude and latitude. The blank node is not a true entity because it does not have an essential identity. Functionally, the solution works, resulting in additional query work to navigate the blank node.

Labeled Property Graph (LPG)

The other main graph type is the Labeled Property Graph. This type of graph is implemented by vendors such as Neo4j, AWS (Neptune), and Azure (Cosmos). Nodes in an LPG are identified by an internal value, not a resource identifier as is the case with triplestores. LPG nodes can have properties that correspond to triples that have literals as objects. One can implement entities by creating a property on a node with a resource identifier as a value. **Figure 5** shows the Leonardo da Vinci entity as an LPG node.

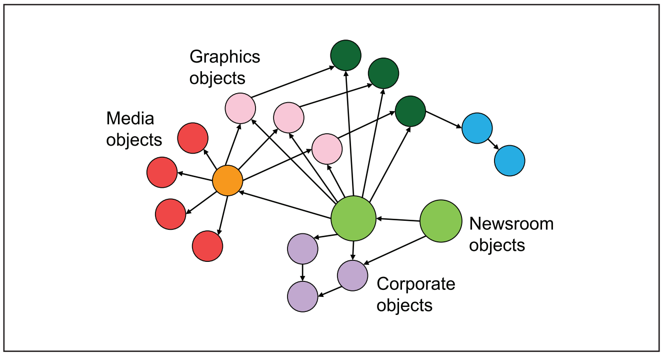


Figure 6. Graph data for news story.

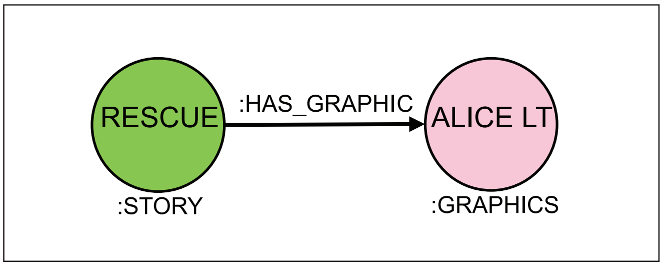


Figure 7. Single triple.

Here, the values are directly associated with the node, not connected via edges. While the property names are like edge labels, they don't inherently carry the same semantic value as an ontological resource. It is possible to approximate ontological IRI naming with convention. For example, neo4j will convert predicates into property names by adding the prefix to the name, e.g., "foaf__name."

Also, LPGs allow for properties on edges, which can make data modeling easier. DEL graphs can achieve the same effect, but they require more complex data model structures than expected.

Finally, LPGs allow one or more labels on nodes and edges that can be interpreted as types.

Graph Queries

Knowledge graphs have a different method for querying data than traditional databases. Instead tables and columns, graph queries work in terms of nodes, edges, and values. The general approach is to look for patterns in the graph. RDF graphs can be queried in the SPARQL language.[14] Property graphs are often queried with the CYPHER[15] or Gremlin[16] languages.

If, for example, we wanted to find things located in the Louvre from the example, a SPARQL query would look something like the following:

```
PREFIX ex: <http://example.org/resources/>
SELECT ?creativeWork WHERE {
 ?creativeWork ex:DisplayedIn ex:Louvre .
}
```

This query returns all node identifiers for subject nodes connected to the object node representing the Louvre via the DisplayedIn relation. Graph query languages can specify complex patterns, filter on literal values, and can test for the existence or non-existence of graph paths.

Example Use Case, Revisited

As mentioned earlier, several data elements were created in a Neo4j LPG to support the hypothetical news scenario in a way that represents a common distribution of data over typical stores. **Figure 6** shows a depiction of a graph that contains entities that reference those data elements and the relationships among them. The visualization is done in the Neo4j Desktop application.

Table 1. Story Properties.

Name	Value
"IRI"	"wgphnews:4657d811-57ca-4be2-8943-a5c370d-c4d4a"
"rdfs__label"	"RIP CURRENT RESCUE"
"wgph__slug"	"RESCUE"

Table 2. Graphic Element Properties.

Name	Value
"IRI"	"wgphgfx:1034189-ab71-4f20-a802-447110ec00ae"
"rdfs__label"	"RIP CURRENT RESCUE Alice Wilson LT"
"wgph__short"	"ALICE LT"

The nodes are depicted in circles such that the color denotes the native data store for the entities.

- Orange, red – audio/visual assets in a content management system
- Light Green – newsroom database objects
- Pink – graphics assets
- Blue – external entities
- Dark Green – entities not in databases but referenced by native data.
- Purple – corporate database

A focus on a smaller section can show the basics of the approach.

This triple shows a news story with a relationship to a graphic overlay element (**Fig. 7**). The node properties for the news story are as follows:

This node is labeled as a STORY with an identifier in the "IRI" property (**Table 1**). This identifier can be used to get the native story data from the appropriate database. The format of the IRI here is not strictly legal but is condensed for brevity. The "wgphnews" namespace of the identifier indicates that the story data can be retrieved from the newsroom database. If the story asset can be accessed from an HTTP URL, then that URL would be a better identifier. The label and slug properties are replicated here to give users a human-readable name for the node.

The story is related to its graphic elements through the edge labeled HAS_GRAPHIC where the story node is the subject, and the graphic element node is the object.

The node properties for the pictured graphic element are shown in **Table 2**.

```
MATCH (story)-[:HAS_EDITED_MEDIA]->(:ebucore__EditorialObject {IRI: "wgphmedia:3ad1c0d1-598b-8ea1-88ee1b8a1de"})
RETURN story.rdfs__label, story.IRI
```

story.rdfs__label	story.IRI
"RIP CURRENT RESCUE"	"wgphnews:4657d811-57ca-4be2-8943-a5c370dc4d4a"

Figure 8. Query to find stories that use a given video asset.

```
MATCH (r)-[:REPORTED_BY]->(s:STORY)-[*1..5]->(fd {IRI: "wgphetc:Shoreville,_New_Hamphire_Fire_Department"})
RETURN DISTINCT r.IRI
```

story.IRI
"wgphstaff:Janet Smith"

Figure 9. Query to find reporters who have interacted with the Shoreville FD.

1545-0279/24©2024SMPTE

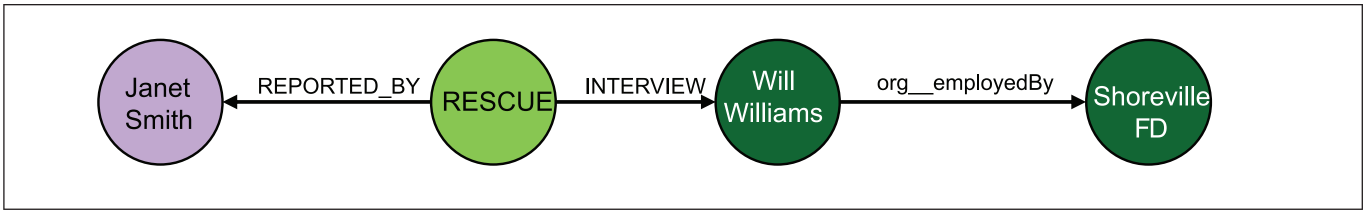


Figure 10. Reporters who have interacted with the Shoreville FD.

This node is labeled GRAPHICS and has an identifier in the "IRI" property. The wgphgfx namespace indicates that the identifier value can be looked up in the graphics database. As with the story node, a human-readable label is set as a property.

The features of this small example can be extended to the full graph above. Nodes represent entities such as people, places, and database records. Edges represent relationships between nodes. Properties contain literal values associated with nodes and edges. Software that understands the node types (encoded here as labels), edge types, and property values can effectively query the graph without using the native databases directly.

This decoupling of the graph from the native databases also has an inverse benefit. In **Table 2**, the story entity in the newsroom database has a relationship with an edited video asset in the content management system. The newsroom software establishes the relationship in its database. If the video entity does not have a corresponding native relationship to the story, then native content database workflows that may benefit from that relationship will not be possible. Once the graph is built, workflows can be based on graph queries instead of native queries to navigate the relationships. It is thus not necessary to modify the content management schema to add the story relationship.

Figure 8 shows a query on the graph that looks for stories that use a given edited package along with the results.

The query input is the IRI of a video asset along with a pattern that follows the "HAS_EDITED_MEDIA" edge back to a referencing story. If there were no connected stories, the result would be null. The value of this function is to find the connection between story assets and media assets without having to combine specific native queries if they even exist.

Figure 9 shows a more complicated query for reporters handling any stories involving the Shoreville (NH) Fire Department.

Notice that the query notation "*1..5" is used – it allows for a graph path of up to 5 hops. The path matching the query is shown in **Fig. 10**.

This subgraph shows the path between Janet Smith and the Shoreville FD through Will Williams, the assistant fire chief. This kind of query would be difficult to execute in traditional databases storing typical data models.

These examples demonstrate how a knowledge graph can be used to aggregates entities and relationships from different types of databases into a single layer. The graph can be queried to use the relationships to build workflows over the entities.

One final benefit of using knowledge graphs is that they can store data enhancements extracted from native data sets by machine learning techniques. ML can find entities and relationships from text, audio, and video. A knowledge graph is a natural fit for organizing and refining these kinds of enhancements. The following section describes the symbiosis between knowledge graphs and ML.

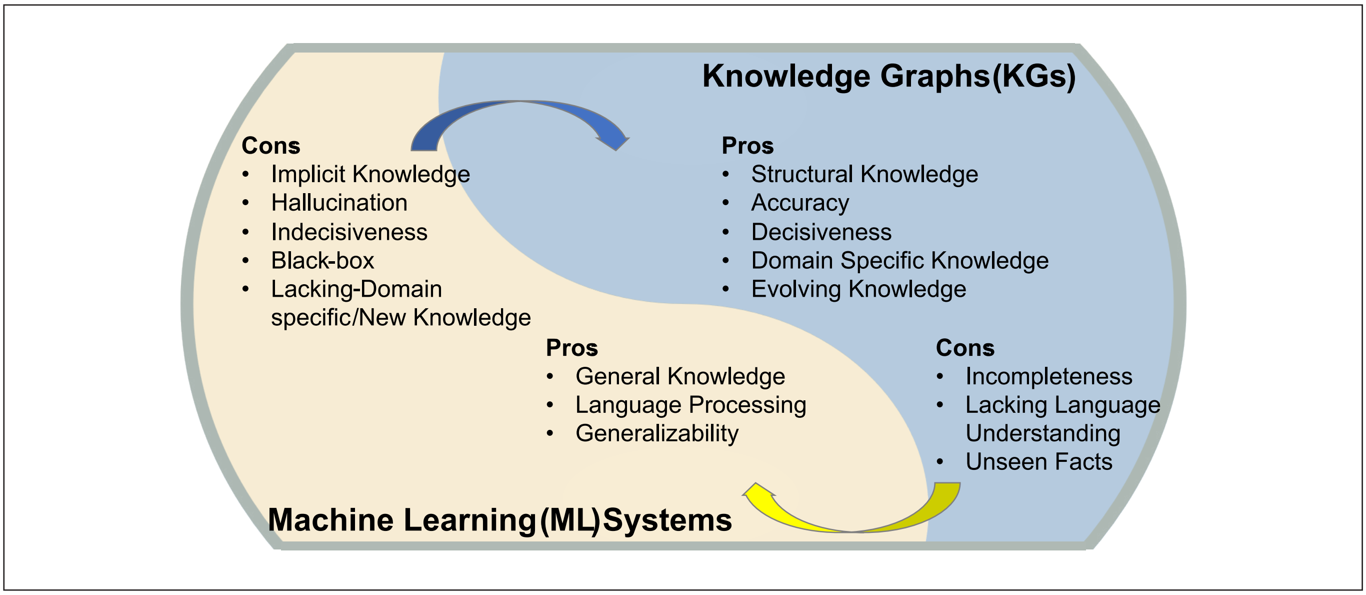


Figure 11. Pros and cons for LLMs and KGs.

1545-0279/24©2024SMPTE

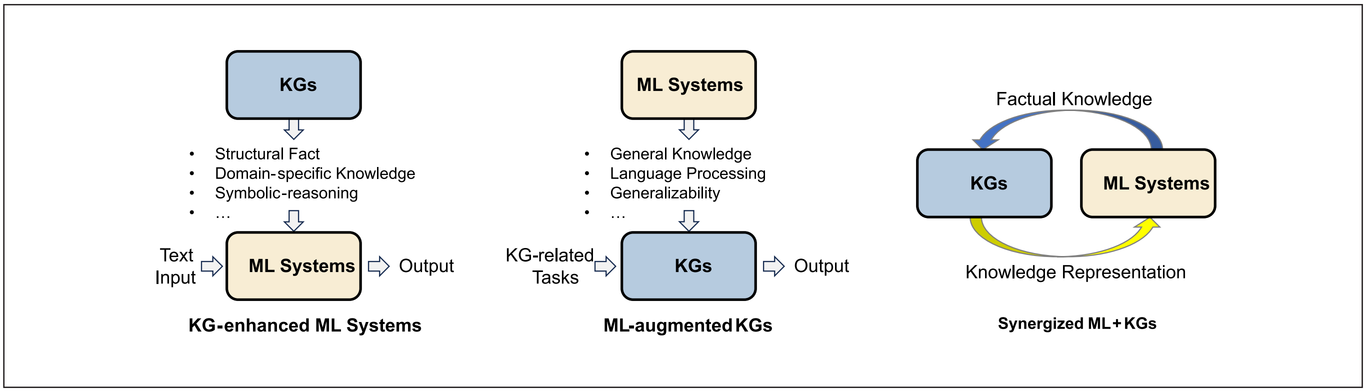


Figure 12. Types of KG and LLM integration. KG-enhanced ML systems (left), ML-augmented KGs (center), synergized ML + KGs.

Integrating Knowledge Graphs and Machine Learning

KGs and ML can enhance recommendation systems' accuracy, interpretability, and personalization. Integrating these technologies leverages the structured, semantic information of KGs with the predictive capabilities of ML.

Background on ML

Machine learning, a branch of artificial intelligence, develops algorithms that learn from data for predictive or decision-making purposes. ML models, applied in domains like natural language processing and recommendation systems, recognize patterns, classify data, and predict outcomes using historical data.

Contrasting Knowledge Graphs with Machine Learning Systems

ML systems, including large language models (LLMs) like ChatGPT[17] and GPT-4,[18] and semantic embedding systems like CLIP,[19] excel in processing textual and visual data but struggle to capture factual knowledge. KGs store factual knowledge effectively but are challenging to build and update. Unifying ML and KGs could merge their strengths, enhancing recommendation systems with external knowledge.[20]

KGs are structured representations of entities and their relationships, offering interpretability and context for recommendation models. However, they often require manual construction and adaptation, making them labor-intensive. **Figure 11**, adapted from Pan et al.,[20] contrasts the pros and cons of KGs and LLMs.

Knowledge Graphs with ML Systems, Better Together

Integrating KGs and ML can leverage both strengths, with ML models learning patterns from KGs, and KGs providing structured context to ML models. This synergy enhances predictive power and interpretability in recommendation systems. **Figure 12**, from Pan et al.,[20] showcases three integration modes: KG-enhanced ML Systems, ML-augmented KGs, and their synergized combination.

Innovative approaches include using LLMs for KG construction and augmentation, as proposed in "LLMs for Knowledge Graph Construction and Reasoning."[21] Another study, "Multi-level Recommendation Reasoning over Knowledge Graphs with Reinforcement Learning,"[22] utilizes knowledge graphs for user interest modeling in recommendation systems. Additionally, "Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs"[23] explores zero-shot recognition using Graph Convolutional Networks and semantic embeddings.

Integrating KGs and ML promises to improve recommendation systems through combined structured information and learning capabilities, enhancing accuracy, interpretability, and personalization.

Acknowledgments

We would like to thank Jennifer Lim and Paul Randall for reviewing our paper and providing feedback.

Conclusion

Knowledge graphs can effectively represent entities and relationships with meaning defined by semantic specification. The structures of these entities and relationships can be processed in a way that provides unique value, especially when entities exist across multiple heterogeneous databases. This aggregate data layer allows workflows to query the graph to find entities based on the relationships and use them to fulfill processing requirements.

Machine learning can extract semantic information from data and use that data to enhance knowledge graphs, increase knowledge about the entities in the system, allow inferences to be made, and give insight into patterns of data organization. Conversely, knowledge graphs can provide context to improve the relevance of semantic data extraction.

The combination of knowledge graph and machine learning technologies can enable deep understanding of interrelated data sets with high confidence and information-rich query results.

References

1. Worldwide Web Consortium (W3C). The Semantic Web Made Easy [Online]. Available: https://www.w3.org/RDF/Metalog/docs/sw-easy
2. Linked Data [Online]. Available: https://en.wikipedia.org/wiki/Linked_data
3. Worldwide Web Consortium (W3C) RDF Primer [Online]. Available: https://www.w3.org/TR/rdf-primer/
4. About MovieLabs: Mission [Online]. Available: https://movielabs.com/about/movielabs-mission/
5. Aidan Hogan, et. al., "Knowledge Graphs," *ACM Computing Surveys 54(4)*:1-37, Jul. 2021.
6. Worldwide Web Consortuim (W3C) RDF 1.2 Concepts and Abstract Syntax, Section

1.1: Graph-based Data Model [Online]. Available: https://www.w3.org/TR/32-concepts/#data-model

7. Internet Engineering Task Force (IETF), RFC 3987, "Internationalized Resource Identifiers (IRIs)," January 2005 [Online]. Available: https://www.ietf.org/rfc/rfc3987.txt
8. Internet Assigned Numbers Authority [Online]. Available: https://www.iana.org/
9. Worldwide Web Consortium (W3C) Media Types Issues for Text RDF Formats [Online]. Available: https://www.w3.org/2008/01/rdf-media-types
10. Worlwide Web Consortium (W3C) Resource Description Framework (RDF) Schema Specification 1.0 [Online]. Available: https://www.w3.org/2001/sw/RDFCore/Schema/20010913/
11. Worldwide Web Consortium (W3C) SKOS Simple Knowledge Organization System [Online]. Available: https://www.w3.org/TR/2009/REC-skos-reference-20090818/
12. Worlwide Web Consortuim (W3C) Web Ontology Language (OWL) [Online]. Available: https://www.w3.org/OWL/
13. Renzo Angles, et. al., "Survey of Graph Database Models," ACM Computing Surveys 40(1) 1:1-39, Feb. 2008.
14. Worldwide Web Consortium (W3C) SPARQL 1.1 Query Language [Online]. Available: https://www.w3.org/TR/151-query/
15. What is openCypher? [Online]. Available: https://opencypher.org/
16. Apache TinkerPop™ [Online]. Available: https://tinkerpop.apache.org/gremlin.html
17. Yiheng Liu, et al., "Summary of CHATGPT/gpt-4 research and perspective towards the future of large language models," arXiv preprint arXiv:2304.01852. May 2023.
18. Anis Koubaa. "GPT-4 vs. GPT-3.5: A concise showdown." Apr. 2023.
19. Alec Radford, et al. "Learning Transferable Visual Models from Natural Language Supervision," *Proc. Intern. Conf. on Mach. Learning* (PMLR), Feb. 2021.
20. S. Pan, et al., "Unifying Large Language Models and Knowledge Graphs: A Roadmap," arXiv preprint arXiv:2306.08302. June 2023.
21. Yuqi Zhu, et al. "LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities." arXiv preprint arXiv:2305.13168, May 2023.
22. Wang, Xiting, et al. "Multi-level Recommendation Reasoning over Knowledge Graphs with Reinforcement Learning." *Proc. of the ACM Web Conf. 2022*, pp. 2098-2018, Apr. 2022.
23. Wang, Xiaolong, Yufei Ye, and Abhinav Gupta. "Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs." *Proc. of the IEEE Conf. on Comp. Vis. and Pattern Recognition*. Mar. 2018.

About the Authors

Roger Sacilotto is a software architect with three decades of expertise working on broadcast and post-production asset management workflow solutions along with metadata standards and interoperability. He joined Avid in 1993.

Shailendra Mathur is vice president and chief architect at Avid, where he oversees technology and architecture. Mathur hold multiple patents and presented papers at numerous conferences, panels, and journals on subjects such as AI and computer vision.

Rob Gonsalves joined Avid as their 15th employee in 1989. He helped develop the Avid Media Composer, specializing in programming video effects, color correction. His work on multi-camera editing software led to Avid winning a Technology & Engineering Emmy Award. He is currently researching the use of AI for media production.

Presented at the SMPTE 2023 Media Technology Summit, Hollywood, CA, 16-19 Oct. 2023. Copyright © 2024 by SMPTE.

DOI: 10.5594/JMI.2024/XBZC2345
Date of publication: 1 April 2024

What's New for Members in 2024!

You asked and we delivered!
Now, included with your membership, you can download any and all SMPTE Standards whenever and wherever you need. We've made all your SMPTE Standards more accessible than ever before!

Access to ALL SMPTE Standards

EVOLVE with us
SMPTE

1545-0279/24©2024SMPTE

TECHNICAL PAPER

KEYWORDS VIDEO COMPRESSION // VIDEO CODEC // MPEG-2 // H.264, AVC // HEVC // VVC // ARTIFICIAL INTELLIGENCE

The Future of Video Compression—Moving Beyond Hybrid Codecs with Machine Learning

By Thomas Guionnet, Marwa Tarchouli, Thomas Burnichon and Mickaël Raulet

Abstract
The consumption of video content on the internet is increasing at a constant pace, along with an increase of video quality. As an answer to the ever-growing demand for high-quality video, compression technology improves steadily. About every decade, a new major video compression standard is issued, decreasing bitrate by a factor of two. Interestingly, the technology does not change radically between codecs generations. Instead, the same principles are re-used and pushed further. There have been several attempts to depart from this model, but none achieved to be competitive. Recently, the research community has started focusing on deep learning-based strategies, with speculation arising as to whether it could be a new contender to the classical approach. This paper analyzes the benefits and limitations of deep learning-based video compression methods and investigates practical aspects such as rate control, delay, memory consumption, and power consumption. Overlapping patch-based end-to-end video compression strategy is proposed to overcome memory limitations.

The consumption of video content on the internet is increasing constantly, along with an increase in video quality. Cisco[1] estimates that by 2023, two-thirds of the installed flat-panel TV sets will be ultrahigh-definition (UHD), up from 33% in 2018. The bitrate for 4K (UHD-1) video is more than double the HD video bitrate and about nine times more than standard-definition (SD) bitrate. As an answer to the ever-growing demand for high-quality video, compression technology is improving steadily. Video compression is a highly competitive and thriving field of research and industrial applications. Billions of people are impacted, from TV viewers and streaming addicts to professionals, from gamers to families. Video compression is used for contribution, broadcasting, streaming, cinema, gaming, video surveillance, social networks, video-conferencing, military, etc.

The video compression field stems from the early 80s. Since then, it has grown with continuous improvements and strong attention from the business side—the video encoder market is planned to reach USD 3.3 Billion by 2027.[2] About every decade, a new major video compression standard halves the required bitrate to achieve a given quality. The latest milestone is the Versatile Video Coding (VVC) standard, issued in 2020. From generation to generation until VVC, coding efficiency has been improved by relying on the same principle: the block-based hybrid video coding scheme.[3] For more than 30 years, the video compression field has known no revolution or disruption. Instead, the same principles and ideas have been re-used and pushed further. At each generation, existing tools are enhanced, and new local coding tools are added, but the overall structure remains the same. In other words, each generation is a complexified version of the previous one. The implementation complexity directly reflects

1545-0279/24©2024SMPTE

the algorithmic complexity increase. For instance, the VVC verification software model (VTM) is about 7 to 8 times slower than its predecessor, the High-Efficiency Video Coding (HEVC) verification model (HM).[4] Many attempts have been made to depart from the block-based hybrid scheme, but few have been successful.

Today, the tremendous progression of video compression technology is not compensating for the increase in the demand for more and higher quality video services. Therefore, the research effort is ongoing, seeking improvements over VVC, as it was over each previous codec generation. Indeed, The Joint Video Expert Team (JVET), a working group managed by both ISO/IEC MPEG and ITU-T VCEG international standardization bodies responsible for the development and support of VVC, is currently conducting explorations beyond VVC. A new situation is arising though: this exploration follows two distinct tracks. One is "classical," consisting of adding or enhancing coding tools to VVC, while the other is dedicated to exploring the usage of machine learning (ML). The field of ML, particularly deep learning (DL), has made dramatic advances during the past decade, especially in the computer vision domain. There are several ways of applying ML to video compression. One can consider creating elementary coding tools, replacing or complementing the existing tools in the hybrid block-based scheme. At the other extremity of the spectrum, it's now possible to completely replace the hybrid block-based scheme with a deep learning model. The latter solution is highly disruptive as it concerns current video compression history, and leads to speculation as to whether ML is destined to become a necessary component in video compression.

At the other extremity of the spectrum, it's now possible to completely replace the hybrid block-based scheme with a deep learning model. The latter solution is highly disruptive as it concerns current video compression history, and leads to speculation as to whether ML is destined to become a necessary component in video compression.

This paper aims to analyze the benefits and limitations of deep learning-based video compression methods, and to investigate practical aspects such as rate control, delay, memory consumption and power consumption. First, the evolution of video compression is recounted, with a few words on previous attempts to depart from the hybrid block-based model. Second, the deep-learning strategies are described, focusing on tool-based, end-to-end, and super-resolution-based strategies. The practical limitations for industrial applications are then analyzed. Finally, a technology is proposed, namely overlapping patch-based end-to-end video compression, to overcome memory consumption limitations. Experimental results are provided and discussed.

A Short History of Video Compression

CODECs and Applications

The idea of temporal prediction for video compression can be tracked back to 1929, with a patent advocating the coding of successive image differences,[5] but the modern history of video compression starts in the 1980s. Two organizations are essentially responsible for video coder/decoder (codec) standardization:[6,7] the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) Video Coding Expert Group (VCEG), a United Nations Organization (formerly CCITT),[8] and the International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) Moving Picture Expert Group (MPEG).[9] Aside from standardization, many proprietary or independent codecs exist. Nonetheless, the most successful and well-known line of codecs stems from standardization and constitutes the focus of this paper.

The first standardized video codec, ITU-T H.120,[10] was issued in 1984 and then updated in 1988. It already includes a form of intra-prediction (Digital Pulse Coded Modulation, DPCM), scalar quantization, entropy coding in the form of variable length coding (VLC), and motion compensation.

ITU-T H.261[11] was first issued in 1988. It is dedicated to video telephony and introduces the most important block-based motion compensation and Discrete Cosine Transform (DCT). It is the first practically successful video codec. It was later replaced by the dramatically improved ITU-T H.263[12] in 1995.

Meanwhile, ISO/IEC developed MPEG-1,[13] which was issued in 1993. It was designed to compress VHS-quality raw video, thus enabling the first digital TV applications (videos CD, Cable, satellite). One may note that the best-known part of MPEG-1 is the MP3 audio format it introduced. MPEG-1 has been followed by MPEG-2/H.262[14] and MPEG-4 part 2,[15] also called MPEG-4 visual, because of its object-oriented approach.

Interestingly, in the 1990s, two lines of standards coexisted. The ITU-T H.26X line was designed for video telephony, while the ISO/IEC MPEG was intended for digital TV broadcasting. However, both shared many technological aspects. Quite logically, ISO/IEC MPEG and ITU-T VCEG have been joining their effort in the development and publication of common video compression standards, thus starting a particularly successful line of video codecs.

MPEG-2/H.262[14] has been a tremendous success in the 1990s, and the enabler of widespread digital TV. MPEG-2 has been present on cable TV, satellite TV, and DVD and is still running nowadays. In the early 2000s, AVC/H.264[16] was a key component of the HD TV development on traditional networks as well as on internet and mobile networks. AVC/H.264 is also used in HD Blu-ray discs. Ten years later, in the 2010s,

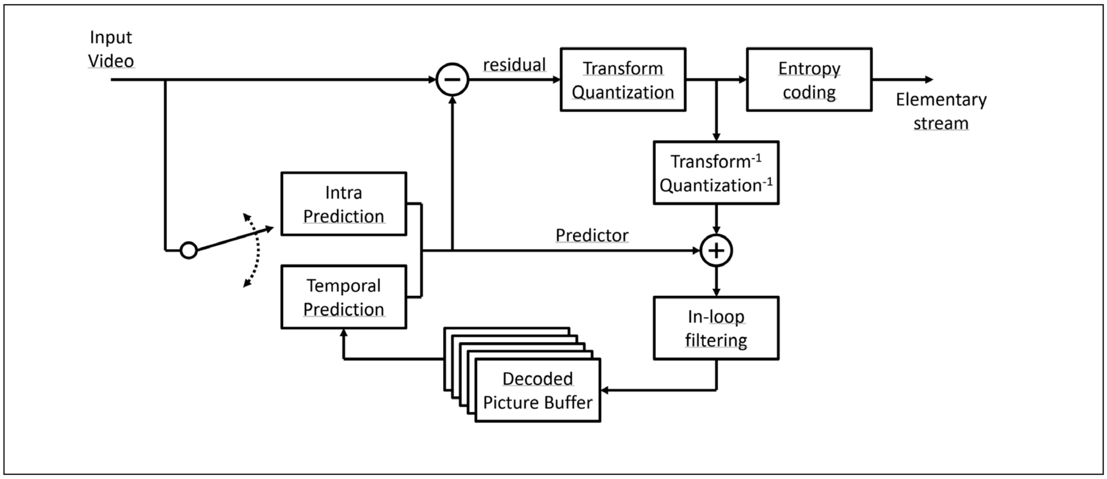


Figure 1. The block-based hybrid video coding scheme.

HEVC (H.265)[17] was the enabler of UHD-1, HDR and WCG. Finally, VVC (H.266)[18] was issued in 2020. Although it is a young codec, not yet widely deployed, it is perceived as an enabler for 8K (UHD-2)[19] and as a strong support for the ever-growing demand for high-quality video over the internet.

The block-based hybrid video coding model is depicted in **Fig. 1** The main elements are:

- Intra-prediction, for coding intra frames, i.e., frames without temporal dependency, or intra blocks inside inter frames, for managing local areas that cannot be temporally predicted.
- Inter-prediction allowing temporal prediction from previously encoded frames.
- Transform for compacting the residual information on a few coefficients.
- Quantization adjusting the trade-off between quality and bitrate.
- Entropy coding, a fundamental information theory concept, determining the statistically shortest representation of the data.
- In-loop filtering, correcting partially coding artifacts, for both better quality and better temporal prediction.

The technology does not change radically between codec generations. Instead, the same principles and ideas are re-used and pushed further. Of course, there are new coding tools, but the overall structure remains the same.

Compared to MPEG-2, AVC/H.264 brought notably reduced complexity integer discrete cosine transform, multiple reference inter-frame prediction, in-loop deblocking filter, variable block sizes, and flexible handling of interlaced video, all contributing to its excellent coding efficiency. During the same period, the video compression research community also focused on the concept of 3D wavelet filtering.[20] The wavelet transform has been used successfully in the JPEG 2000 image compression standard.[21] The wavelet transform is a signal decomposition and analysis tool. Applied to an image, it replaces usual transforms such as DCT for energy compaction and provides a resolution-scalable representation. A wavelet compressed image can be reconstructed progressively, from lowest to highest frequencies, without coding efficiency loss. When applied to video, the same principle is extended to a 3D pixel volume.[22] The MC-EZBC codec is a good example of state-of-the-art 3D wavelet-based video coding.[23] This kind of technology was promising but has yet to reach the AVC/H.264 performance.[24]

Jumping to the next generation, HEVC brought many improvements over AVC/H.264, including a much more flexible partitioning scheme, with up to 64 x 64 pixels partition sizes instead of 16 x 16, multiple transforms, improved motion compensation filtering, and a new loop filtering restoration tool, Sample Adaptive Offset (SAO).

In parallel, a strong research focus was set on sparse modeling for image and video representation and analysis. As explained in,[25] sparse coding represents data with linear combinations of a few dictionary elements. Generally, the dictionary must be learned in order to be best suited to the data. Considering images, the underlying idea is that only a tiny subset of the huge set of all possible pixel value combinations represents viewable images. Therefore, images can be represented by smaller variables, as few as possible if compact representation is desired. Although the idea seems quite simple, building a dictionary is a non-trivial task.[26,27] These methods have yet to reach the general performance and flexibility of HEVC. One may note, however, that sparse coding

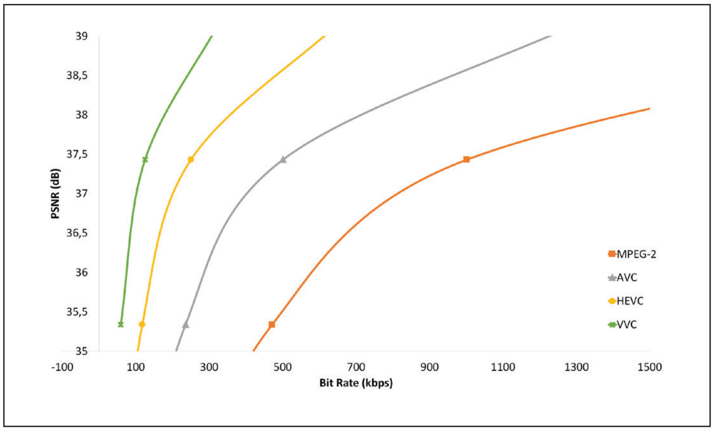


Figure 2. Video codecs rate distortion performance progression example.

shines in specialized applications, such as very low bitrate human face coding.[25] Also, the dictionary learning strategy anticipates the upcoming machine learning.

Finally, VVC outperforms HEVC thanks to enhanced coding tools, such as an even more flexible partitioning scheme or a new in-loop restoration filter. Moreover, VVC includes from start several features that make it "versatile," including 360° video coding, screen content coding, gradual decoder refresh for low delay applications, and scalability based on reference picture resampling (RPR), the ability to perform temporal prediction on reference images of different resolutions.

> TODAY, THE **RECOGNIZED INDUSTRY BENCHMARK** FOR VIDEO COMPRESSION PERFORMANCE IS VVC, AND THE QUESTION MAY BE RAISED AS TO THE POSSIBILITY OF EXCEEDING VVC PERFORMANCE.

Each codec generation allows decreasing the bitrate approximately by a factor of two (**Fig. 2**). This comes, however, at the cost of increased complexity. For instance, the reference VVC encoder is about 7 to 8 times more complex than the reference HEVC encoder.[4]

Today, the recognized industry benchmark for video compression performance is VVC, and the question may be raised as to the possibility of exceeding VVC performance. All indications point toward this, and the JVET standardization group is currently conducting explorations. The Ad-Hoc Group 12 (AHG12) is dedicated to enhacing VVC. Around 15% Bjøntegaard Delta Rates[28] (BDR) gains were already achieved only two years after VVC finalization,[29] so it's likely that this process may continue for at least another decade.

However, there is a new contender—artificial intelligence, or more precisely, machine learning or deep learning. In another ad-hoc group, AHG11, JVET is exploring how machine learning can be the basis of new coding tools. This also brings BDR gains of about 12%.[30] This leads to speculation about machine learning's role in video compression's future. At this stage, two facts should be pointed out.

First, in considering the "traditional" methods explored by AHG12, a coding tool seems to stop bringing gains: frame partitioning. The partitioning is a fundamental tool for video compression. It specifies how precisely the encoder can adapt to the unique characteristics of local content. The more flexible it is, the better the coding efficiency. All the subsequent coding tools depend on the ability to partition the frame efficiently. During the exploration following HEVC standardization, enhancing the partitioning brought up to 15% BDR gains.

Similarly, in the AHG12 context, people came up with new extended partitioning strategies. However, only marginal gains were reported.[31] This could mean there may be a limit to further improvement.

The second fact is the development of end-to-end deep-learning video compression. This strategy is highly disruptive. In short, the whole block-based hybrid coding scheme is replaced by a set of deep learning networks, such as auto-encoders. These schemes now compete with state-of-the-art fixed image and video coders.[32,33] This level of performance was reached in just five years. That's an unprecedently fast progression. Even if the progression slows down, one may easily extrapolate that state-of-the-art video compression performance will soon be the end-to-end strategy prerogative. Therefore, a turning point in the video codec history may have been reached.

Machine Learning-Based Video Compression

Tool-Based ML-Based Video Compression

As AI and ML are gaining attention in all possible research fields, video compression is no exception. The JVET standardization group has started an exploration activity dedicated to introducing ML in the VVC framework.[34] The idea here is to keep the hybrid block-based model and to replace or complement elementary coding tools with ML-based tools.

Intra-prediction is addressed in Refs. 35 and 36. Both approaches consider the prediction of a block of pixels from a causal pixel neighborhood. The prediction is performed by a neural network (NN) replacing the traditional directional or planar intra-prediction. The NN is expected to be able to predict complex shapes and textures. 2 to 3% Bjøntegaard Delta Rates (BDR) gains are reported.

Inter-prediction is considered using several strategies. Galpin et al., proposed an enhanced bi-prediction mode.[37] Instead of predicting a block with an average of two motion-compensated reference blocks, the two predictors are fed to an NN, which outputs the final prediction. Up to 1% BDR gains are reported. Ma et al. considered a similar approach but with a single predictor.[38] In Ref. 39, a whole reference frame is generated by a NN. This new frame is added to the reference list for temporal prediction. Therefore, each image block can be predicted either from a past encoded frame or from an NN-generated frame without fundamentally changing the encoding/decoding process. Up to 2% BDR gains are reported.

A strong focus has also been set on loop filtering.[34] The general idea of loop filtering is to restore already encoded frames. It serves both as a post-processing, improving the visual quality of the compressed video, and as a BDR improvement, as it allows better temporal predictions. The first attempts considered replacing all the loop filters (LF) with an ML process. The idea has then been refined with adaptive methods trying to take advantage of the best of two worlds: signal processing and ML. The CNNLF[40] is proposed as an alternative to the deblocking and SAO filters of VVC. It is up to the encoder to decide locally whether to activate CNNLF. The filter inputs data, including quantization parameter (QP), a prediction image, and a partition image. The filter also includes a scaling as a post-processing after the NN step. Up to 12% BDR gains are reported, illustrating significant impact of the single loop filter coding tool. The filter proposed Ref. 41 replaces all VVC LF, though it can be turned off at block level. It similarly relies on rich input and post-scaling while making use of attention models. Similar performance is reported.

Overall, the coding efficiency gains obtained with these approaches are largely significant. However, they come at an unprecedented cost in complexity, with figures going up to 400 times slow-down of the decoder.[40,41]

Super-Resolution-Based Video Compression

The idea of super-resolution stems from the well-known over-the-top (OTT) streaming concept. Depending on the available bitrate, an optimal combination of resolution and compression tuning exists. In other words, when the bitrate decreases, it becomes more efficient to reduce the video resolution rather than getting more compression artifacts. Considering the fact that ML has been studied as a means of recovering fine details when increasing image resolution, one arrives naturally to the idea that encoding videos at a lower resolution may lead to a better trade-off than the current approaches, thanks to the capability of NN to up-sample content without generating the traditional blurring and aliasing phenomenon.

A traditional codec decomposes the video sequence before being compressed using a traditional codec. It is then synthesized to retrieve the original resolution.[42] The decomposition consists of down-sampling only the inter-coded frames. Thus, the intra-frames carry texture information, while inter-frames carry temporal information. Synthesis, or up-sampling, is assisted by a motion- compensated NN. Up to 9% BDR gains are reported.

The video is encoded at a lower resolution.[43] NN-based super-resolution is applied as a post-processing to recover the original resolution. However, to better adapt to frame characteristics, the last layer of the NN is specialized for each sequence. The corresponding parameters are transmitted along with the video stream. About 6.5% of coding gains are reported.

Complexity gains can be observed depending on the complexity of the chosen NN approach. Indeed, thanks to the lower video resolution, the coding/decoding step is much less complex, potentially compensating for the NN complexity.[44]

End-to-End Learned Video Compression

Nowadays, learned image and video compression has been the target theme for both the machine learning and image/video compression communities. In this context, the Challenge on Learned Image Compression (CLIC)[45] aims to encourage both communities to advance the image and video compression field using machine learning algorithms by either designing new codec architectures or introducing new perceptual metrics.

Inspired by the success of learned image compression, where the state-of-the-art outperforms the latest traditional coding system VVC,[46,47] video coding attracted more focus from the research community.

Learned video compression approaches can be divided into two main categories. The first one regroups generic models based mainly on the autoencoder architecture and trained on a large dataset. For instance, Lu et al., introduced the first low-latency compression framework called Deep Video Compression (DVC) following the traditional coding pipeline while using auto-encoders to code motion vectors and residuals, a pretrained optical flow model for motion estimation and a bilinear warping for motion compensation.[48] The method used in Ref. 49 improves the DVC performance by using multiple frames as references. This new coding system is called MLVC. Four new neural modules were added. They aim to investigate past frames to reduce temporal redundancies more effectively. In the same context of low-latency coding,[50] introduces a recurrent learned video codec (RLVC) using a recurrent autoencoder and a recurrent probability model to compress the motion and the residual features. While DVC manages to outperform the low-delay P frames (LDP) configuration of x264[51] and compete with the same configuration of x265,[52] MLVC and RLVC outperform DVC and x265[52] in terms of coding efficiency. Hierarchical Learned Video Compression (HLVC), described in Ref. 53 presents a framework to code a group of pictures (GOP) structure, which includes P and B frames, with different levels of quality. P and B frames are coded with hierarchical quality levels using neural models based on CNNs, auto-encoders, and RNNs. This work's proposed framework depends on GOP structure, which is set manually before proceeding to the training stage. Although this method codes a GOP structure with B and P frames, it was evaluated against the LDP mode of x265[52] and the low latency model DVC. Compared to x265, it gains Bjøntegaard Delta Bit Rate (BDR) of -6% for PSNR models and −35.94% for MS-SSIM models. A neural coding framework dedicated to I and P frames is proposed.[54] An architecture adapted to I, P, and B frames is presented in Ref. 55. It contains two networks: MOFNET is dedicated to motion estimation and compensation, while CodecNet performs residual coding. This approach competes with the state-of-the-art video codec HEVC (HM 16.20). Li et al., explored conditional coding in their proposed contextual video compression framework, DCVC.[56] This work was further improved by enhancing temporal context in the coding system,[57] while Li et al.[33] focused on improving the entropy model by including spatio-temporal information in the parameters estimation step.[57] and[33] achieved, respectively, a BD-rate (PSNR oriented) gain of -14.4% against HEVC (HM) and -4.7% against VVC (VTM13).

Other approaches exploit deep learning techniques and algorithms in learned video compression, such as 3D convolutions,[58] Generative Adversarial Networks (GANS),[59] and transformers.[60]

The second category of learned video codecs focuses on approaches based on lightweight neural models that can be adjusted to the content of each video sequence to be encoded. The idea of implicit neural representations (INR) is to adjust the coding model to the content of each sequence while choosing a light model architecture. The encoding step includes training the neural codec on the sequence to be coded; the bitstream is formed by quantifying and entropy coding the weights of the model. In the decoding process, the weights of the neural codec are retrieved, and inference is performed to reconstruct the source input, which is also the training sequence. Dupont et al., has first leveraged the neural network architecture to compress implicitly a video frame, using a simple fully connected network (FCN), composed of fully connected layer and non-linear activation functions, that maps the spatial pixel coordinates (x, y) to their RGB values.[61] Kim et al., adapted the previous method to inter-coding by using two models, the first one, non-neural model, maps the spatial and temporal pixel coordinates (x, y, t) to a latent representation and the second one is an FCN model which maps the latent to the RGB values.[62] The intra-coding model[63] went one step further by relying on two models as well as an overfitted latent representation. The first model serves as an entropy model that helps entropy decoding the latent representation and then the second model reconstruct the RGB pixel values. Those methods are called pixel-wised implicit representations, which means that the neural model reconstructs the sequence frames pixel by pixel, thus making the process very slow. Chen et al. proposed an image-wise model which recovers a whole frame from only the temporal index.[64] They change the architecture of the neural codec by adding a convolutional block to the FCN network. This method improves the encoding speed by 25x to 70x and the decoding speed by 38x to 132x comparing to pixel-wise approaches while improving the video quality. Other works try to improve the performance of Neural Representations for Video (NeRV) either by adding the spatial information as input to the network,[65,66] or by including the content information,[67,68] or by enhancing the temporal correlation between frames.[69,70] In terms of coding efficiency, FFNeRV[70] seems to achieve significant gains (1db for the same rate) compared to the baseline NeRV.[64] In terms of coding speed, HNeRV[67] improve the coding speed of E-NeRV,[65] which is already faster than the baseline NeRV,[64] by x16.

All in all, although auto-encoder based codecs managed to outperform the latest handcrafted codec VVC in terms of coding efficiency, they are large and hefty systems very demanding in terms of hardware. Furthermore, their performance depends on the training dataset. In contrast, the content-adaptation approaches are not constrained by those limitations; they have yet to reach a level of maturity that allows them to compete with VVC. Their state-of-the-art currently competes with HEVC. Nevertheless, considering the rapid pace of advancement in this field, surpassing H.264 in 2021, outperforming HEVC in 2022 and currently exceeding VVC, one could predict that the content-adaptation methods can have a significant progress in a short period of time.

Practical Application of Machine Learning Based Video Compression

Performance Evaluation

With the emergence and the fast progress of end-to-end learned video codecs, a question arises about a methodology for comparing these models with traditional codecs, given the difference in their design. Several factors may bias such comparisons.

First, certain learned video codecs presented in the state of the art are not evaluated in practical test conditions, unlike handcrafted codecs. For example, the bitstream is not always generated in learned compression methods, and the decoding time might be incredibly long.

Handcrafted video codecs are designed to process YUV (luminance and chrominance) video sequences while most learned video codecs are trained on RGB clips. Therefore, one should be vigilant on the choice of format on which quality metrics are computed, in order to prevent any bias towards one approach over the other.

Furthermore, the results are dependent on the metric used. Many end-to-end methods[53,54] are reported to surpass significantly traditional codecs when MS-SSIM is used as objective quality metric. However, one must note that in this context, end-to-end methods are optimized for MS-SSIM, through the inclusion of the metric in the loss function, while traditional codecs are optimized for PSNR. Therefore, such comparison tends to favor end-to-end methods.

Finally, when comparing a learned video codec with a traditional one, it is important to consider the configuration and the GOP structure. For example, HLVC[53] employs a fixed GOP structure that corresponds to random access with P and B frames using a GOP size of ten frames. In contrast, traditional codecs are optimized for larger GOP sizes (32 or 64 frames). Thus, evaluating HLVC[53] against a traditional codec such as HEVC puts automatically HEVC either at an advantage, if the default HEVC configuration is used, or at a disadvantage, if the HEVC GOP are aligned on HLVC.[53]

Nonetheless, recent works[55,71] on learned compression put their approaches in practical test conditions, which made their comparison to handcrafted video codecs more relevant. Moreover, in the same context, the challenge CLIC[45] provides an evaluation framework where a bitstream should be written, and maximum model size is fixed as well as a method to assess quality.

Delay, Rate-Control and Content Adaptation

There is a huge difference between a codec, as defined by standards, and a ready for production live video encoder. The codec is only a part of a video encoder. A video encoder must

1545-0279/24©2024SMPTE

manage various inputs for capture, decoding, encoding, muxing and output, along with system functions and user interface. Even when focusing on the encoding part, there is more than the codec. Live encoding requires optimization of the complexity/quality trade-off, which generally translates into added delay. This delay must of course stay under control. Delay is caused among other things by pre-processing and analysis in a look-ahead buffer, frame reordering for efficient group of pictures (GOP) structure coding, pipelining, and rate-control.

Content adaptation is desirable for optimal quality. GOP structure is generally adapted to the nature of the content. Scene-cuts are detected, and temporal prediction is avoided across them. Considering end-to-end video coding, the same ideas may apply. However, depending on the end-to-end implementation, it may be simpler. The idea of GOP structure may be managed in a transparent manner by the ML model. The notion of successive GOP may be easily conserved, allowing easy chunking for OTT and short zapping time. One may note that INR, such as NeRV-like methods is inherently structured as GOP from arbitrary sizes.

In short, content adaptation does not seem to be an obstacle to the end-to-end video encoders development. Rate-control, on the other hand, may be more difficult. Indeed, in traditional video coding, there is an understandable, though non-trivial, relationship between the QP and the bitrate. In an end-to-end video encoder, the control of the bitrate depends on the strategy considered. For an auto-encoder, there exists a parameter tuning the bitrate. However, the effect of this parameter is generally not easy to model. Some encoders are trained for a single value of this parameter. It implies that if one needs 64 rate levels, like the 64 QP values of VVC, 64 models must be trained and stored. To address this issue, the methods used in[72] proposes a new loss function, where the λ parameter, responsible for rate tuning, is not constant. It allows the design of a training procedure where several values of λ are fed randomly to the system, thus making a model that can react appropriately to any value of λ at inference time. For content adapted end-to-end video codecs, the bitrate is related to the size of the model, hence the need to design several models, or some kind of scalability in the structure. Literature on this topic is limited as of today, and there is no doubt that further research is needed.

Finally, the main difficulty to handle may very much be the huge operational complexity of NN.

Computing Resources

During several years after the deep learning emergence, researchers did not really care about models' complexity. The solutions proposed for various public challenges were more complex every year, while their performances continued to grow exponentially.

The case of learned image and video compression models is no exception, particulary generic models based on the autoencoder architecture. Actually, even though the field of end-to-end neural compression improves promptly, yielding promising results, the complexity of the proposed models increases also significantly. In fact, the training as well as the inference process of such models can be consuming in terms of time, and demanding in terms of computing resources. For example,[60] proposed a neural video compression model based on transformers and autoencoders, demonstrating a BDR gain against HEVC. However, its heavy and complex architecture makes its replication extremely difficult. While most neural compression method tends to overlook informations about their model complexity, the number of parameters is typically around 20 million parameters, if not more. Moreover, their training process can take days, sometimes months depending on the available hardware resources.

> CONTENT ADAPTATION **DOES NOT SEEM TO BE AN OBSTACLE** TO THE END-TO-END VIDEO ENCODERS DEVELOPMENT. RATE-CONTROL, ON THE OTHER HAND, MAY BE MORE DIFFICULT.

The increase in the model's complexity has been made possible thanks to the hardware evolution. For deep learning technologies, Graphical Processing Units (GPUs) are often the default choice, because of their ability to perform many low-level mathematical operations in parallel. Initially designed for games and graphically intensive applications, researchers thought their capabilities were suited to run deep learning models. This market is dominated by Nvidia, and since the deep learning development, they have built new GPU architectures that make their hardware more effective for model training and inference. But this kind of hardware still is a general-purpose solution. Some manufacturers decided to build specific chips designed to run deep learning models even more effectively. One can think about Google Tensor Processing Units (TPUs), or Microsoft Catapult project. They are based respectively on Application-Specific Integrated Circuits (ASIC) and Field Programmable Gate Array (FPGA) and allow power consumption reduction related to GPUs. These solutions are available in cloud infrastructures, so they can be used for models training and online inference. These use cases are rarely constrained by consumption resources. If more power is needed to speed up training or inference, it is simple to scale by adding GPUs for example; however, “edge devices” need to be considered.

These are appliances on which data collection takes place. It can be desktop computers, smartphones, or connected devices. While GPUs or TPUs are still the default solutions for training models, a lot of work has been done for performing inference on edge devices. Contrary to cloud platforms, scaling is very hard due to space, power, and connectivity limits. But this is a very important use case as it allows processing data locally, mitigating network limitations, increasing security, and improving data privacy. Researchers and manufacturers have put a lot of effort into improving edge computing hardware for processing deep learning models. Hence, new types of AI-optimized accelerators have been designed during the past few years, that can be regrouped under the name Neural Processing Units (NPUs). Main mobile manufacturers have designed their own solution. This includes chips such as the Apple Neural Engine, the Kirin 980 from Huawei, or the Exynos 9820 from Samsung. There also exists development boards such as the Nvidia Jetson Nano or the Google Coral Edge TPU. NPUs are based on specific architectures that make deep learning model execution faster while having limited consumption. A lot of accelerators exist today,[73] and this is a very active research field. A few years ago, MLPerf benchmarks[74] was released in order to make AI platform performance comparisons simpler. It provides training time, inference time, and more recently power consumption of a specific hardware configuration for different AI models. Despite these initiatives, AI accelerators comparison remains very hard as performances are related to too many factors, not only the accelerator itself. Performances are also impacted by the CPU, and the software library used to deploy the model.

In addition to work on specialized hardware, a lot of work has been done on the software part. Some of them are designed for CPUs (Basic Linear Algebra Subprograms) OpenBLAS, Intel Math Kernel Library (MKL), ...), and others for GPUs (cuBLAS, cuDNN (Deep Neural Network), ...). All of them optimize matrix operations in order to make AI model execution faster using only algorithmic optimizations. These are libraries allowing low-level mathematic operations, but they are mainly used through higher-level frameworks and tools. For example, Openvino[75] and TensorRT,[76] respectively developed by Intel and Nvidia, are platforms offering runtimes with optimized operations implementation, but also some model optimization strategies. This includes weights quantization, network pruning, or operations fusion.

The combination of hardware and software evolution allows the execution of powerful AI models on edge devices in real-time. But the AI field is evolving really fast. As models become more complex, hardware and software providers must improve their solutions to speed up and reduce energy use. Recent trends such as neuromorphic computing[77] show there is room for improvement with completely different designs. Also, new hardware is challenging dominant existing solutions. For example, the Hailo 8 chip[78] presents performances up to 13x those of Google TPUs. All of this shows that the Moore’s law continues and makes possible further improvements in AI.

A Case-Study: End-to-End Memory Consumption

Problem Statement

As models’ sizes are growing continuously, memory consumption is also becoming an issue, along with computing power. The case of end-to-end learned encoding is considered here. The auto-encoder architecture, built with convolutional layers, enables processing different video resolutions, no matter the resolution used during the training step. However, with growing models’ sizes and video resolutions 4K and 8K (UHD-1 and UHD-2), these solutions are facing hardware memory saturation. One way to solve this issue is to use a patch-based coding approach. The video frames are divided into patches smaller than the frame size, which can be encoded independently. Then, the decoded patches are gathered to reconstruct the decoded frames.

This solution addresses the hardware limitation issues, but the reconstructed frames can have block artifacts at the patch boundaries, widely deteriorating the video quality.

Patch-Based End-to-End Video Encoding

Our proposed solution to the previous problem, as described in Ref. 79, involves performing patch encoding while simultaneously removing block artifacts. The input frame is divided into overlapping patches, horizontally and vertically, which are fed to the neural compression model. At the output of this model, the decoded patches are recovered and gathered to reconstruct the decoded frame. To remove block artifacts and reconstruct overlapping areas, a weighted linear function is used to blend reconstructed pixels.

This method is evaluated on JVET Common Test Conditions (CTC) sequences (8-bit sequences),[80] using our implementation of the neural codec described in Ref. 32, while varying the number of overlapping pixel noted N from 0 to 32. Then, BDR gains are computed compared to the full image coding using the same neural encoding model. This model was trained to optimize MSE as a distortion metric.

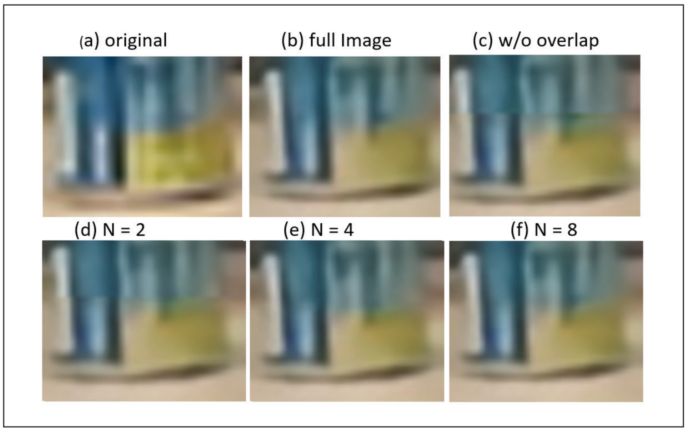


Figure 3. Visual results of comparison for FourPeople. The model used optimizes the MSE metric with λ = 4096.

1545-0279/24©2024SMPTE

Table 1. Performance of patch based end-to-end encoding.

Resolution	Method	Coding Time GPU 2080 11GB	Coding Time GPU 3090 24GB
1920x1080	Full Resolution Coding	OOM	OOM
	Patch coding in parallel with overlapping	3.82s	2.05s
1280x720	Full Resolution Coding	OOM	0.93s
	Patch coding in parallel with overlapping	1.91s	1.012s
832x480	Full Resolution Coding	1.06s	0.52s
	Patch coding in parallel with overlapping	1.10s	0.55s

Without overlapping, our approach shows a loss in BDR (Average BDR +0.013), corresponding to the block artifacts caused by patch coding. However, as the number of overlapping pixels increases, increasing BDR gains are observed. This is an unexpected benefit of the method, as the goal was to be neutral in terms of coding efficiency. With N = 2, the BD-rate gain among CTC sequences is -0.025 proving that two overlapping pixels are efficient to remove border artifacts. For N = 8, BDR gains increase to : -0.034 and with N = 16, a saturation of these gains is observed.

The model used optimizes the MSE metric with λ = 4096. In **Fig. 3,** the visual results of the proposed approach is illustrated on the Four People sequence, using different sizes of overlapping pixels. As a conclusion, the border artifacts are attenuated when N = 2 and N = 4, and they are completely removed when N= 8.

The experimental complexity of this method is measured by computing the the coding time and the memory consumption (**Table 1**). To generate these results, two powerful GPUs: a GeForce RTX 2080ti and a GeForce RTX 3090 with memory capacities of 11 Gbytes and 24 Gbytes respectively, were used.

Coding a 1920x1080 frame turns out to be impossible for both GPUs, while coding an 1280 x 720 frame can only be achieved with the GPU possessing the higher memory capacity (24 Gbytes). On the other hand, for frames with smaller resolutions, our approach adds more complexity (about 3%) due the additional overlapping pixels processing.

To conclude this section, the proposed approach addresses the hardware memory limitation problem, while maintaining same or better quality as the full resolution learned coding.

Conclusion

From MPEG-2 in the 90s to VVC in use today, four successive major generations of codecs have made video ubiquitous, from TV screen to smartphones, from over-the-air to internet. All these codecs are based on the same general structure, the hybrid block-based model. Previous attempts to overcome this model have all failed, despite their numerous technical qualities and features, leading to speculation as to whether this model will continue to dominate.

Today, a small hint of a decline of the hybrid block-based model is being observed, along with the rise of machine learning. Machine learning is the state-of-the-art technology in many image and video processing fields, but still not in video compression. While not yet suitable for video compression, ML is making rapid progress. The argument is made in this paper that current limitations can be addressed, either through plain technological progress, or through dedicated algorithmic progress.

As an example, a new method of memory management for ML-based end-to-end image and video compression is described in this paper, namely patch encoding with overlapping. Also, the recent evolution of implicit neural representations, with more reasonable decoding complexity, is encouraging.

All in all, for the upcoming video codec generation, two approaches are competing. Time will tell, but our guess is that there will be another generation of the hybrid block-based model before the advent of ML-based video compression. Researchers just need a few years to refine and make the technology practical. Model sizes and hardware capabilities will eventually converge.

References

1. Cisco Annual Internet Report (2018–2023) White Paper. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html
2. Video Encoder Market with COVID-19 Impact by Number of Channel (Single, Multichannel), Mounting Type (Standalone, Rack-mounted), Application (Broadcast, Surveillance (Commercial, Residential, Institutional)), and Geography - Global Forecast to 2025. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/video-encoder-market-109133493.html
3. W. Thomas and S. Heiko, "Video Coding: Part II of Fundamentals of Source and Video Coding," *Foundations and Trends® in Signal Processing*, 10 (1–3): 1-346, Jan. 2016 [Online]. Available: http://dx.doi.org/10.1561/2000000078,.
4. F. Bossen, X. Li and K. Suehrin, "AHG report: Test model software development (AHG3)," JVET-T0003-v1, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 20th Meeting, teleconference, 7-16 Oct. 2020.
5. R. D. Kell., "Improvements relating to electric picture transmission systems, British Patent Patent 341,811, 1929
6. G. Sullivan, "Overview of International Video Coding Standards (preceding H.264/AVC)," ITU-T VICA Workshop, 22-23 July 2005, ITU Headquarter, Geneva, 2005
7. S. V. K. S. G. ATHISHA, "An Overview Of H.26x Series And Its Applications," *International J of Eng Science and Techno*, 2(9): 4622-4631, 2010.
8. International Telecommunication Union (ITU) [Online]. Available: https://www.itu.int
9. Motion Picture Experts Group (MPEG) [Online]. Available: https://www.mpeg.org
10. "H.120 : Codecs for videoconferencing using primary digital group transmission," ITU-T Recommendation H.120 (11/88) [Online]. Available: https://www.itu.int/rec/T-REC-H.120-198811-S/en
11. "Video codec for audiovisual services at p x 64 kbit/s," ITU-T Recommendation H.261 (03/93). [Online]. Available: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=1088&lang=en
12. "Video coding for low bit rate communication," ITU-T Recommendation H.263 (01/05). [Online]. Available: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=7497&lang=en
13. "MPEG-1 (Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s)," ISO/IEC JTC 1/SC 29, ISO/IEC 11172-x:1993.
14. "Information technology—Generic coding of moving pictures and associated audio information: Video," ITU-T Recommendation H.262 (02/12). [Online]. Available: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11627&lang=en
15. International Organization for Standardization (ISO). "ISO/IEC 14496-2:2004: Information technology—Coding of audio-visual objects—Part 2: Visual.
16. "Advanced video coding for generic audiovisual services," ITU-T Recommendation

1545-0279/24©2024SMPTE

H.264 (08/21). [Online]. Available: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=14659&lang=en.
17. "High efficiency video coding," ITU-T Recommendation H.265 (08/21). [Online]. Available:https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=14660&lang=en.
18. "Versatile video coding," ITU-T Recommendation H.266 (04/22) [Online]. Availabe: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=14948&lang=en.
19. T. Biatek, M. Abdoli, T. Guionnet, A. Nasrallah and M. Raulet, "Future MPEG standards VVC and EVC: 8K broadcast enabler." [Online]. Available: Available at: https://www.ibc.org/technical-papers/future-mpeg-standards-vvc-and-evc-8k-broadcast-enabler/6754.article, 14 September 2020.
20. E. Moyano, F. Quiles, A. Garrido, T. Orozco-Barbosa and J. Duato, "Efficient 3D wavelet transform decomposition for video compression," pp. 118-125, 2001, DOI: 10.1109/DCV.2001.929950.
21. D. S. Taubman, M. W. Marcelin and M. Rabani, "JPEG2000 Image Compression Fundamentals, Standards and Practice," Springer Science & Business Media, vol. 642, 2012.
22. V. Bottreau, M. Benetiere, B. Felts and B. Pesquet-Popescu, "A fully scalable 3D subband video codec," *Proc. 2001 Intern Conf. on Image Process*. 7-10 Oct. 2001.
23. P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," IEEE Trans. Circuit Syst. Video Technol., 14 (10):1183-1194, Oct. 2004.
24. P. Lambert, W. P. De Neve, I. Moerman, P. Demeester and R. Van de Walle, "Rate-distortion performance of H.264/AVC compared to state-of-the-art video codecs," Jan. 2006.
25. J. Mairal, F. Bach and J. Ponce, "Sparse Modeling for Image and Vision Processing," 2014. DOI: https://doi.org/10.48550/arXiv.1411.3230.
26. Y. Sun, T. Xu, T. X. M. and J. Lu, "Online dictionary learning based intra-frame video coding via sparse representation," 2012.
27. M. Mahdavi-Nasab and H. Irannejad, "Block Matching Video Compression Based on Sparse Representation and Dictionary Learning," 2018.
28. G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," *Technical Report VCEG-M33*, ITU-T SG16/Q6, TX 2001.
29. JVET-Z0012-v1, "JVET AHG report: Enhanced compression beyond VVC capability (AHG12)," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 26th Meeting, by teleconference, 20–29 April 2022.
30. JVET-Z0023, "EE1: Summary of Exploration Experiments on Neural Network-based Video Coding," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 26th Meeting, by teleconference, 20–29 April 2022.
31. JVET-Y0150-v2, "EE2-1.1: Tests on unsymmetric partitioning methods," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 25th Meeting, by teleconference, 12–21 Jan. 2022.
32. Z. Cheng, H. Sun, M. Takeuchi and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," 2020.
33. J. Li, B. Li and Y. Lu, "Hybrid spatial-temporal entropy modelling for neural video compression," *Proc. of the 30th ACM Internat. Conf. on Multimedia*, pp 1503-1511, 2022.
34. Elena Alshina et al., "JVET AHG report: Neural network-based video coding," JVET-AA0011-v1, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 27th Meeting, teleconference, 13–22 July 2022.
35. M. Meyer and C. Rohlfing, "AHG11-related: Investigation on CNN-based Intra Prediction," JVET-U0105-v3, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 21st Meeting, teleconference, 6–15 Jan. 2021.
36. T. Dumas et al., "EE1 test 3.1: intra prediction using neural networks," JVET-Y0082-v2, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 25th Meeting, teleconference, 12–21 Jan. 2022.
37. F. Galpin et al., "AHG11: Deep-learning based inter prediction blending", JVET-V0076-v2," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 22nd Meeting, teleconference, 20–28 Apr. 2021.
38. Changyue Ma et al., "AHG11: Neural Network Based Motion Compensation Enhancement for Video Coding," JVET-Y0090-v1," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 25th Meeting, teleconference, 12–21 Jan. 2022.
39. Zizheng Liu et al., "AHG11: NN-based Reference Frame Interpolation for VVC Hierarchical Coding Structure, JVET-Y0096-v1," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 25th Meeting, teleconference, 12–19 Jan. 2022.
40. Liqiang Wang et al., "Neural Network Based in-Loop Filter with Constrained Memory," *Proc. 2022 IEEE Internat. Conf. on Multimedia and Expo (ICME)*, 2022, DOI: 10.1109/ICME52920.2022.9859910.
41. Yue Li et al., "EE1-1.2: Test on Deep In-Loop Filter with Adaptive Parameter Selection and Residual Scaling, JVET-Y0143-v2," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 25th Meeting, teleconference, 12–21 Jan. 2022.
42. Ming Lu et al., "EE1: Tests on Decomposition, Compression, Synthesis (DCS)-based Technology, JVET-V0149," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 22th Meeting, teleconference, 20–28 Apr. 2021.
43. Takeshi Chujoh et al., "EE1.2: Additional experimental results of NN-based super resolution (JVET-U0053)," JVET-V0073-v1, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 22nd Meeting, teleconference, 20–28 Apr. 2021.
44. Chaoyi Lin et al., "EE1-2.3: CNN-based Super Resolution for Video Coding Using Decoded Information, JVET-Y0069-v21," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 25th Meeting, teleconference, 12–21 January 2022 DOI: 10.1109/VCIP53242.2021.9675417.
45. Workshop and Challenge on Learned Image Compression (CLIC). [Online]. Available: https://www.compression.cc/
46. G.-H. Wang, J. Li, B. Li and Y. Lu, "EVC: Towards Real-Time Neural Image Compression with Mask Decay," The Eleventh International Conference on Learning Representations (ICLR), May 2023.
47. D. He, Z. Yang, W. Peng, R. Ma, H. Qin and Y. Wang, "ELIC: Efficient Learned Image Compression With Unevenly Grouped Space-Channel Contextual Adaptive Coding," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5718-5727, 2022.
48. G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai and Z. Gao, "DVC: An End-To-End Deep Video Compression Framework," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11006-11015, 2019.
49. L. Jianping, L. Dong, L. Houqiang and W. Feng, "M-lvc: : Multiple frames prediction for learned video compression," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
50. R. Yang, F. Mentzer, L. V. Gool and R. Timofte, "Learning for Video Compression With Recurrent Auto-Encoder and Recurrent Probability Model," *IEEE Journal of Selected Topics in Signal Processing*, 15 (2): 388-401, 2021.
51. VideoLAN. [Online]. Available: https://www.videolan.org/developers/x264.html
52. VideoLAN. [Online]. Available: https://www.videolan.org/developers/x265.html
53. R. Yang, F. Mentzer, L. V. Gool and R. Timofte, "Learning for Video Compression With Hierarchical Quality and Recurrent Enhancement," *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6628-6637, 2020.
54. H. Liu, M. Lu, Z. Ma, F. Wang, Z. Xie, X. Cao and Y. Wang, "Neural Video Coding Using Multiscale Motion Compensation and Spatiotemporal Context Model," *IEEE Transactions on Circuits and Systems for Video Technology*, 31: 3182-3196, 2021.
55. T. Ladune, P. Philippe, W. Hamidouche, L. Zhang and O. Deforges, "Conditional Coding for Flexible Learned Video Compression," *Neural Compression Workshop ICLR 2021*, 2021.
56. J. Li, B. Li and Y. Lu, "Deep Contextual Video Compression," *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.
57. X. Sheng, J. Li, B. Li, L. Li, L. Dong and L. Yan, "Temporal Context Mining for Learned Video Compression," *IEEE Transactions on Multimedia*, 25: 7311-7322, 2023, DOI: 10.1109/TMM.2022.3220421.
58. A. Habibian, T. van Rozendaal, J. M. Tomczak and T. S. Cohen, "Video Compression With Rate-Distortion Autoencoders," Proc. of the *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7033-7042, 2019.
59. S. Santurkar, D. Budden and N. Shavit, "Generative Compression," *2018 Picture Coding Symposium (PCS)*, 2018.
60. F. Mentzer, G. Toderici, D. Minnen, S.-J. Hwang, S. Caelles, M. Lucic and E. Agustsson, "VCT: A Video Compression Transformer," *NeurIPS'22*, 2022.
61. E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh and A. Doucet, "COIN: COmpression with Implicit Neural representations," *Computer Vision and Pattern Recognition*, 2021.
62. S. Kim, S. Yu, J. Lee and J. Shin, "Scalable Neural Video Representations with Learnable Positional Features," *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.
63. T. Ladune, P. Philippe, F. Henry, G. Clare and T. Leguay, "COOL-CHIC: Coordinate-based Low Complexity Hierarchical Image Codec," *arXiv preprint*, 2022.
64. H. Chen, B. He, H. Wang, Y. Ren, L. S. Nam and A. Shrivastava, "NeRV: Neural Representations for Videos," *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.
65. Z. Li, M. Wang, H. Pi, K. Xu, J. Mei and Y. Liu, "E-NeRV: Expedite Neural Video Representation with Disentangled Spatial-Temporal Context," *Proc. European Conference on Computer Vision*, pp. 267–284, 2022.
66. Y. Bai, C. Dong and C. Wang, "PS-NeRV: Patch-wise Stylized Neural Representations for Videos," *arXiv preprint*, 2022.
67. H. Chen, M. Gwilliam, S.-N. Lim and A. Shrivastava, "HNeRV: A Hybrid Neural Representation for Videos," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10270-10279, 2023.
68. H. Chen, M. Gwilliam, B. He, S.-N. Lim and A. Shrivastava, "CNeRV: Content-adaptive Neural Representation for Visual Data," *BMVC 2022*, 2022.
69. Q. Zhao, M. S. Asif and Z. Ma, "DNeRV: Modeling Inherent Dynamics via Difference Neural Representation for Videos," *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2031-2040, 2023.
70. J. C. Lee, D. Rho, J. H. Ko and E. Park, "FFNeRV: Flow-Guided Frame-Wise Neural Representations for Videos," *arXiv preprint*, 2022.
71. O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle and L. Bourdev, "ELF-VC: Efficient Learned Flexible-Rate Video Coding," *Proc. of the IEEE/CVF Intern. Conf. on Comp. Vision (ICCV)*, pp. 14479-1448, 2021.
72. Chaoyi Lin et al., "AHG11: Variable rate end-to-end image compression," JVET-U0102," Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, 21st Meeting, teleconference, 6–15 Jan. 2021.
73. A. Reuther, P. Michaleas, M. Jones, V. Gadepally and J. K. Siddharth Samsi, "AI and ML Accelerator Survey and Trends," *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, 2022.
74. V. J. Reddi, C. Cheng, D. Kanter, P. Mattson and Y. Z. Guenther Schmuelling,

"MLPerf Inference Benchmark," *ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020.

75. OpenVINO. [Online]. Available: https://docs.openvino.ai/latest/index.html#
76. NVIDIA. [Online]. Available: https://developer.nvidia.com/tensorrt
77. N. Imam and T. A. Cleland, "Rapid online learning and robust recall in a neuromorphic olfactory circuit," *Nature Machine Intelligence*, pp. 181–191, 2020.
78. Hailo-8. [Online]. Available: https://hailo.ai/products/hailo-8/
79. M. Tarchouli, S. Pelurson, T. Guionnet, W. Hamidouche, M. Outtas and O. Deforges, "Patch-Based Image Coding with End-To-End Learned Codec using Overlapping.," *12th International Conference on Artificial Intelligence, Soft Computing and Applications*, pp. 53-63, 2022, DOI 10.5121/csit.2022.122305.
80. F. Bossen, J. Boyce, X. Li, V. Seregin and K. Sühring, "VTM Common Test Conditions and Software Reference Configurations for SDR Video," document JVET-T2010 of JVET, Oct. 2020.

About the Authors

Thomas Guionnet, a fellow research engineer at Ateme, heads the Innovation team's research on artificial intelligence for video compression. He has contributed to standardization procedures, taught video compression, and written various publications such as patents, international conference papers, and journal papers.

Marwa Tarchouli received an engineering diploma in electronic engineering from Ecole Nationale Supérieure d'Electronique, Informatique, Télécommunications, Mathématique et Mécanique de Bordeaux (ENSEIRB MATMECA), Bordeaux, France, in 2020. Since 2021, she is a PhD student at Ateme and INSA Rennes.

Thomas Burnichon is ATEME's Director of Technology and focuses on the video transcoding workflows for Live, VOD and OTT streaming applications. He keeps in close contact with content and service providers all over the world to identify best practices and help drive innovations.

Mickaël Raulet is the chief technology officer at ATEME, where he drives research and innovation with various collaborative research and development projects. He represents ATEME in several standardization bodies. He is the author of numerous patents and more than 100 conference papers and journal scientific articles.

Presented at the SMPTE 2023 Media Technology Summit, 16-19 October 2023.
Copyright © 2024 by SMPTE.

DOI: 10.5594/JMI.2024/IPYX8877
Date of publication: 1 April 2024

Standards Technology Committee Meetings

On a quarterly basis, the Standards Community convenes for week-long TC Meetings. During these sessions, participants provide updates on progress and collaborate on advancing standards work.

3-5 June 2024
OTTAWA, CA

18-20 Sept. 2024
GENEVA, CH

Interested in hosting a TC Meeting?

SMPTE

TUTORIAL

KEYWORDS ARTIFICIAL INTELLIGENCE // MACHINE LEARNING // AUTOMATIC SPEECH RECOGNITION

Automatic Speech Recognition with Machine Learning:

Techniques and Evaluation of Current Tools

By Randy Fayan, Zahra Montajabi, and Rob Gonsalves

This research offers an in-depth review of current Automatic Speech Recognition (ASR) methods and their significant impact on media production. It compares the accuracy and performance of top ASR models and services. The study examines key ASR aspects, including voice activity detection, language identification, and multi-language support, and evaluates their accuracy metrics.

1545-0279/24©2024SMPTE

Abstract

This research offers an in-depth review of current Automatic Speech Recognition (ASR) methods and their significant impact on media production, with a focus on the transformer model's self-attention mechanism for understanding sequential relationships. It compares accuracy and performance of top ASR models like Meta's Multilingual Machine Speech, OpenAI's Whisper, and Google's Universal Speech Model along with services from Microsoft Azure, Amazon Web Services, and Google Cloud Platform. The study examines key ASR aspects, including voice activity detection, language identification, and multilanguage support, and evaluates their accuracy metrics. Challenges such as limited data for certain languages and complexities in linguistic nuances are highlighted. Additionally, the paper discusses ASR's role in media production, from creating time-based captions to transforming editing techniques. By analyzing the ASR process from audio preprocessing to post-processing, the research bridges academic and practical perspectives, enabling media producers to utilize advanced ASR technologies effectively.

This investigation aims to provide a comprehensive understanding of the most widely used ASR techniques, with an emphasis on best practices in contemporary media production.

We explore the underlying mechanisms of the new machine learning (ML) models, particularly the crucial role of the transformer model. It provides a background on the transformer model's architecture, focusing on how its self-attention mechanism facilitates superior performance in tasks requiring an understanding of time-based relationships in data, such as in Automatic Speech Recognition (ASR). This background is crucial for comprehending why these models excel in transcription tasks and provides context to their comparative analysis. Through this investigation, we seek to bridge the gap between theory and practical application in the ever-evolving realm of ASR technologies.

We present an in-depth comparison of state-of-the-art ASR approaches, including three open-source models: Whisper from OpenAI,[1] Multilingual Machine Speech (MMS) from Meta,[2] and Universal Speech Model (USM) from Google.[3] We also compare these models with commercial offerings including the transcription services available in Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP). The Experiments section towards the end of the paper summarizes the results of these comparative tests. These methods are evaluated based on their ability to deliver features like transcription, voice activity detection, language detection, speaker identification, and support for multiple languages. Additionally, we measure their performance using metrics such as Word Error Rate, Character Error Rate, and compute time.

In assessing each ASR model's effectiveness for hundreds of languages, notable challenges emerge. The limited data availability for certain languages is a significant hurdle for training robust models. Moreover, the linguistic complexity involving varied dialects, accents, and the complexities of word and sentence segmentation, particularly in continuous speech, present significant difficulties for multilingual transcription.

ASR Applications

Automatic speech recognition is an increasingly crucial tool in media production. The ability to obtain high-quality transcriptions and word timings from recorded audio, and the applications derived from these transcriptions, have cemented ASR's importance in the industry.

For example, automatically generated transcriptions can be used for time-based captioning. ASR-produced transcripts include word-based temporal information, making it possible to overlay the time-based text on video.

Another key application is searching media libraries for specific spoken content. For instance, searching a media library for the phrase "bicycling up a hill" would identify clips and time offsets where this phrase was found. This search can be both exact and fuzzy. A fuzzy search might also identify phrases like "biking up a hill."

ASR has become an invaluable tool for enhancing the production process in scripted shows. The emerging concept of "paper cut" editing leverages ASR transcriptions, allowing editors to craft content on paper before making tangible edits. Additionally, ASR supports automated story assembly where AI systems tap into the transcript to streamline and refine content, ensuring the narrative remains cohesive. While many shows are scripted beforehand, there are times, especially in interviews and

1545-0279/24©2024SMPTE

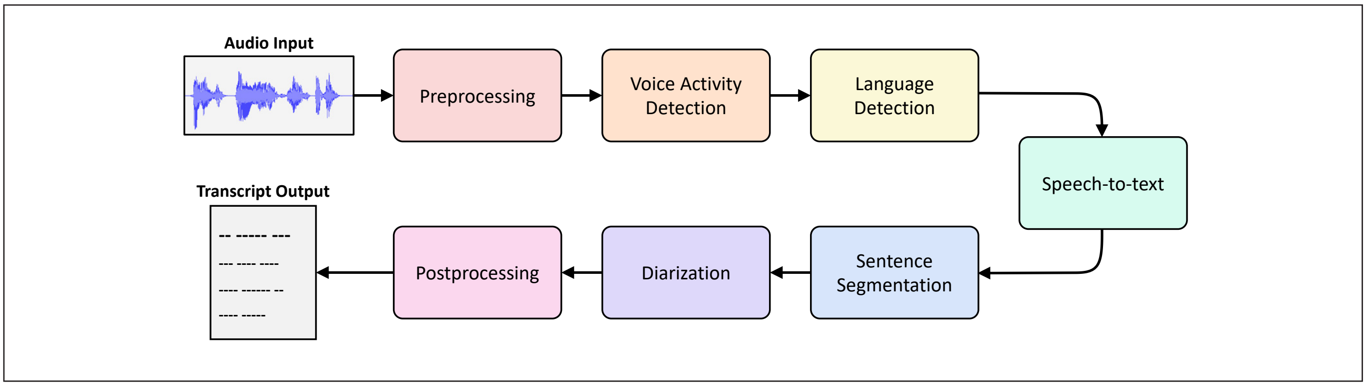


Figure 1. Processes flow for ASR.

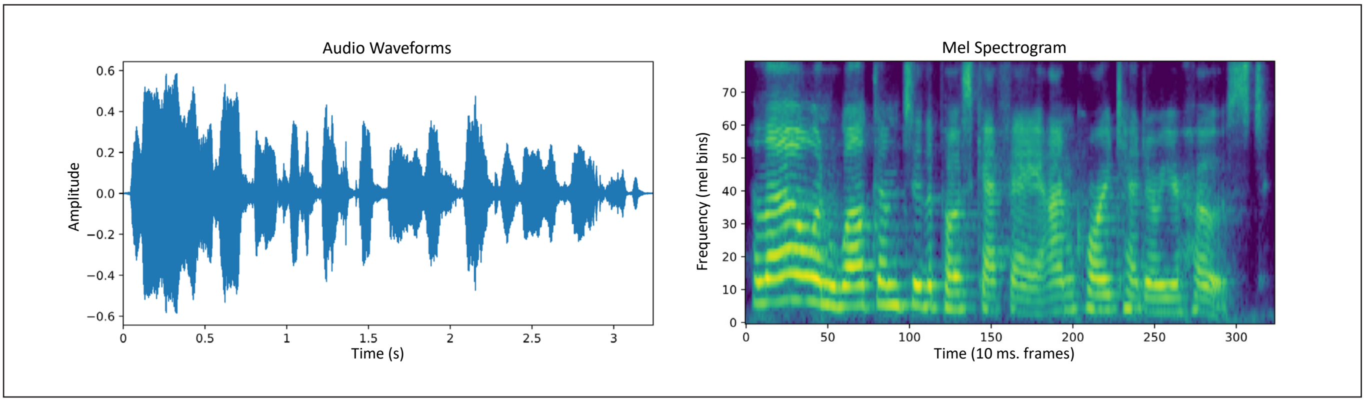


Figure 2. Audio waveform (left), and corresponding mel spectrogram (right).

reality TV, where ASR-generated transcripts become the backbone for script development. Moreover, with a script ready, ASR makes it possible to synchronize media with the script, guaranteeing a seamless alignment of visual content with the scripted narrative.

ASR Processes

ASR is the mechanism of transforming human speech into textual representations.[4] The ASR process steps, as shown in **Fig. 1**, commences with audio preprocessing, where raw audio transforms to the frequency domain to better represent the nuances of speech. Voice activity detection (VAD) then discerns periods of speech from non-speech segments, optimizing the transcription process. The system then identifies the spoken language, a crucial step, especially in multilingual contexts. The speech-to-text conversion, the heart of ASR, translates the spoken content into text while ensuring metadata accuracy. Sentence segmentation further dissects this continuous text stream into discernible sentences, aiding in understanding. Speaker diarization, another critical component, identifies individual speakers in multi-speaker scenarios, providing context to conversations. Finally, post-processing refines the raw ASR output, converting it into standardized formats suitable for various applications, particularly media production.

Each process is described in more detail in the following sections.

Audio Preprocessing

Audio preprocessing is an essential step in ASR, ensuring that the raw audio data is transformed into a format that can be more readily analyzed for speech content. Initially, the raw audio, which is in the time domain, undergoes a transformation to the frequency domain. This is accomplished by first resampling the audio to a 16 kHz rate, an optimal rate to effectively capture the nuances of human speech. A spectrogram of the audio showing the frequency content over time is produced using the "mel" scale. The term mel scale is based on an abbreviation of the word "melody."[5] It simulates human auditory perception.

To understand the mel spectrogram, consider the audio's transformation through the Short-Time Fourier Transform (STFT).[6] The STFT segments the audio into short frames, often 25 msec long, moving in 10-msec strides, to analyze the frequency content within each frame. After this step, a filter bank is applied. This bank consists of 80 triangular windows, which broaden in width with increasing frequency. The uniqueness of these windows lies in their alignment with the mel scale. When this filter bank is applied to the power spectrum derived from the STFT, the result is the mel spectrogram. In essence, the mel spectrogram condenses the information, representing 16,000 original audio samples with 8,000 spectral values per sec, as shown in **Fig. 2**.

The graph on the left shows an audio waveform of a sound file of a person talking. This waveform shows a visual representation of the sound's amplitude variations

1545-0279/24©2024SMPTE

over time, with the x-axis denoting time in seconds and the y-axis denoting the amplitude of the audio signal. Large absolute values in the waveform indicate moments of high energy or loudness; values closer to zero denote quieter portions. The graph on the right illustrates the mel spectrogram of the same audio file. This shows how the frequency content of a signal changes over time but with frequencies mapped to the perceptual mel scale. In this graph, the x-axis represents time divided into frames of 10 msec each, while the y-axis depicts the mel frequency bins. The intensity of colors in the spectrogram indicates the strength or loudness of frequencies at different points in time. This transformation serves as a bridge, converting the intricate patterns in human speech into a format that ASR systems can decipher with high precision.

Voice Activity Detection

Voice activity detection (VAD) is a technique that detects the presence or absence of human speech. Common open-source implementations of VAD systems are WebRTC by Google[7] and the ML-based Silero VAD.[8] Note that the latter operates on mel spectrograms. In general, VAD systems detect the magnitude of speech in an audio stream and decide where the speech segments start and end.[9] For example, **Fig. 3** shows the waveform of an audio clip that has music and silence before people start speaking.

You can see how the VAD system skips over the music and the silence (both indicated in red) and finds the segment where speech occurs (indicated in green).

VAD systems can help the process of transcription in ASR systems by reducing the execution time since the system will not try to transcribe non-speech parts. The VAD process typically runs about ten times faster than the transcription process. Also, using a VAD can help improve the accuracy of overall transcription by preventing the ASR system from mistakenly transcribing non-speech parts of the audio.

Language Detection

The challenge of identifying the spoken language within an audio segment is known as spoken language detection or identification. This capability is important in numerous applications in the media production industry, such as automating the translation and subtitling of multilingual films, facilitating the editing and organization of large-scale video archives, and assisting in creating multilingual broadcasts.

Humans can usually recognize languages just by hearing short clips of speech. However, traditional ASR systems that focus on specific speech features, like rhythm or tone, often struggle with shorter audio samples.[10] A promising solution to this challenge lies in deep learning methodologies. These techniques forgo the conventional phoneme recognition layer and instead delve into a more extensive exploration of feature spaces, enabling the discovery of purely discriminative features tailored for the task. The adoption of these deep learning strategies offers the potential to significantly enhance the accuracy and robustness of spoken language detection systems.

Of the three open-source models we tested for this paper, only OpenAI's Whisper will optionally perform language detection as part of the speech-to-text process. During the

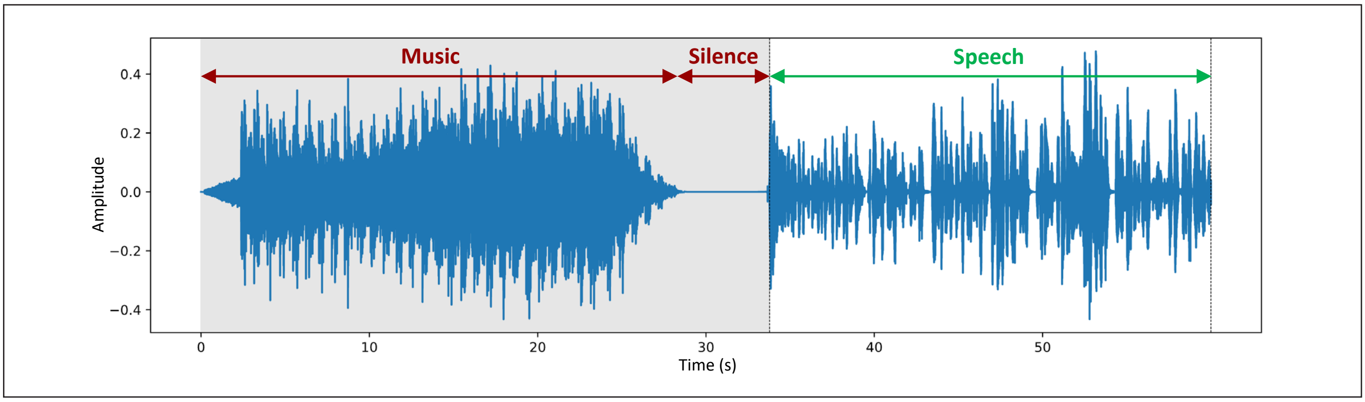
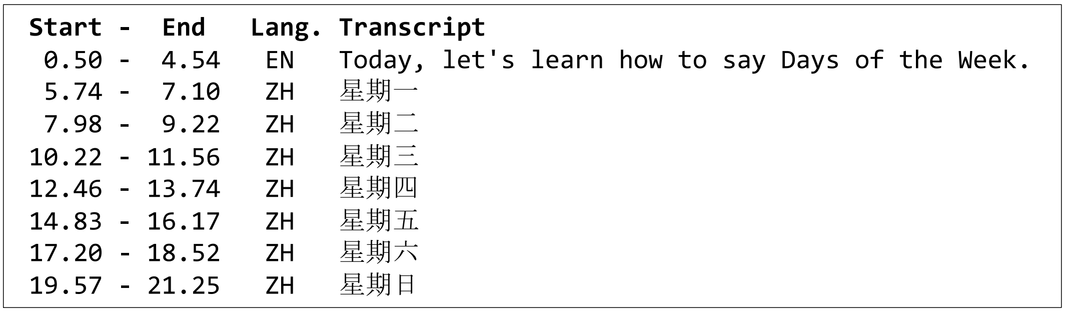


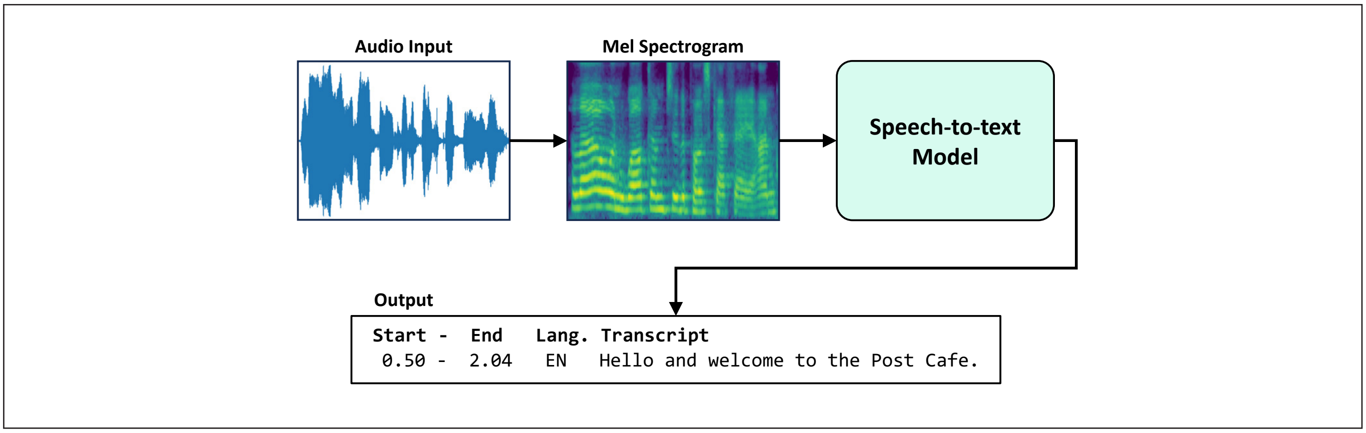
Figure 3. Voice activity detection.



```
Start -   End   Lang. Transcript
 0.50 -  4.54   EN    Today, let's learn how to say Days of the Week.
 5.74 -  7.10   ZH    星期一
 7.98 -  9.22   ZH    星期二
10.22 - 11.56   ZH    星期三
12.46 - 13.74   ZH    星期四
14.83 - 16.17   ZH    星期五
17.20 - 18.52   ZH    星期六
19.57 - 21.25   ZH    星期日
```

Figure 4. Example output of an ASR system with language detection.

1545-0279/24©2024SMPTE



```
Start -  End   Lang. Transcript
 0.50 -  2.04   EN   Hello and welcome to the Post Cafe.
```

Figure 5. Components of a speech-to-text system.

training process for Whisper, the expected output transcriptions were labeled with the spoken language in the input file, so the system learned to predict the language when transcribing audio files, as shown in the example in **Fig. 4**. Meta's MMS has a separate model for language detection, and Google's USM doesn't do this at all.

Speech-to-text

Speech-to-text technology enables machines to convert spoken language into written text. It is a crucial component of ASR systems, acting as the interface between the spoken audio input and the textual output, as shown in **Fig. 5**.

ASR systems aim to transcribe the spoken content as accurately as possible and produce metadata like word/phrase start-times, end-times, and the speaker's language. These tasks are complex due to the variability in speech signals caused by factors such as accents, speaking rate, background noise, and speaker characteristics. Recent advancements in ML and AI have significantly improved the performance of these systems.

Pre-training, a concept in the field of ML, has proven to be a game-changer in developing robust ASR systems. Pre-training refers to training a model on a large-scale dataset before fine-tuning it on a specific task.

For instance, in the study "Robust Speech Recognition via Large-Scale Weak Supervision" by Radford et al.,[1] the authors pre-trained their Whisper models on a vast dataset of 680,000 hours of audio transcripts from the internet. This pre-training stage allowed the models to learn high-quality speech representation, significantly improving their performance when fine-tuned on speech-to-text tasks.

Moreover, the study demonstrated that pre-trained models could generalize well to different tasks and environments, often outperforming models trained only on specific tasks. This is a significant finding as it suggests that large-scale, weakly supervised pre-training can effectively improve the robustness and usefulness of ASR systems.

Another significant development in ASR systems is the capabillity to handle multiple languages and tasks. The study "Scaling Speech Technology to 1,000+ Languages" by Pratap et al.[2] presented a single model that can perform ASR across more than 1,000 languages. This was achieved by pre-training the model on a large unlabeled multilingual dataset and then fine-tuning it on a smaller labeled dataset.

In addition to multilingual capabilities, the model supports multitask learning. It can perform a wide range of tasks, including transcription, translation, voice activity detection, and language identification. This is a significant advancement as it allows a single model to replace multiple stages of a traditional speech-processing pipeline.

Scalability is a critical factor in the development and deployment of ASR systems. The ability to scale up the model to handle large volumes of data and a wide range of tasks is key to achieving high performance and broad applicability.

The study "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages" by Zhang et al. introduced the Universal Speech Model (USM), a single large model that performs ASR across more than 1,000 languages.[3] The model was trained on a large dataset of 12 million hours of multilingual audio data and then fine-tuned on a smaller dataset. The study demonstrated that a single large model can effectively utilize large datasets and deliver state-of-the-art performance on multiple tasks.

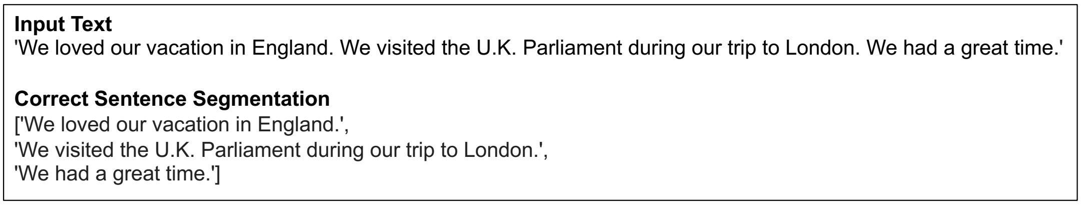
Speech-to-text technology plays a pivotal role in ASR systems. Advancements in ML and AI have significantly improved the performance of these systems. Pre-training, multilingual and multitask training, and scalability are key factors contributing to these advancements. However, the lack of training data due to limited corpora remains challenging for under-represented dialects and languages.[2]

Sentence Segmentation

Sentence segmentation, or sentence tokenization, is a fundamental aspect of Natural Language Processing (NLP) and plays a crucial role in ASR systems. It involves identifying the boundaries of sentences within a continuous stream of spoken or written text. This process is essential for understanding the context and semantics of speech, as each sentence often conveys a distinct idea or concept.

In English and many other languages, punctuation marks such as periods, question marks, and exclamation points typically denote the end of a sentence. However, the task of sentence segmentation is not as straightforward. Punctuation marks are not exclusive to sentence boundaries; they also

1545-0279/24©2024SMPTE



Input Text
'We loved our vacation in England. We visited the U.K. Parliament during our trip to London. We had a great time.'

Correct Sentence Segmentation
['We loved our vacation in England.',
'We visited the U.K. Parliament during our trip to London.',
'We had a great time.']

Figure 6. Example output of sentence segmentation system.

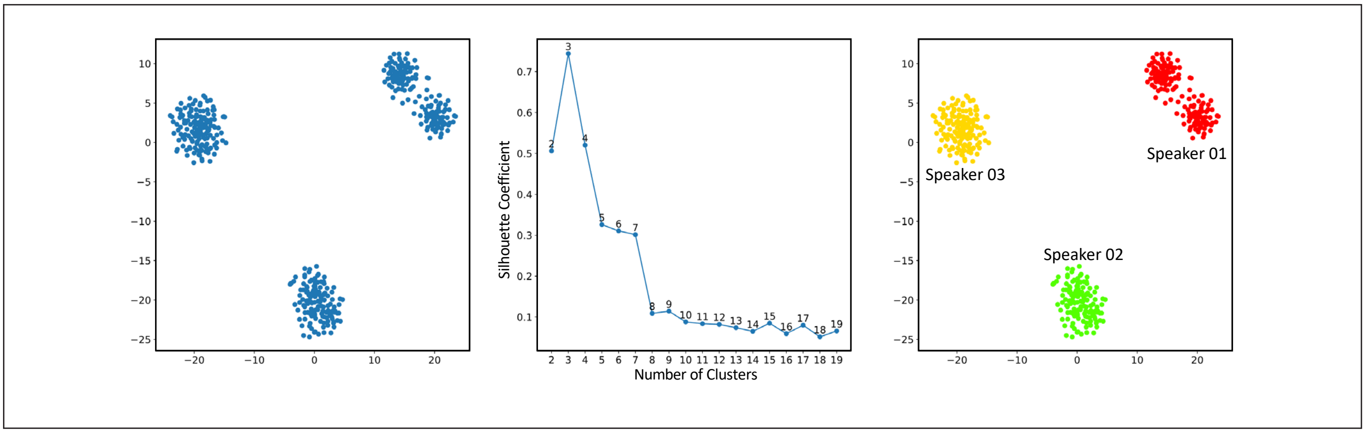


Figure 7. 2D projection of voiceprints (left), silhouette coefficients (middle), clustered voiceprints (right).

appear in abbreviations, acronyms, and other contexts. For instance, the abbreviation "U.K." contains two periods that do not signify the end of a sentence, as shown in **Fig. 6**.

Moreover, in languages like Thai, which lack punctuation, sentence segmentation becomes even more challenging.[11] Ambiguities in punctuation and the lack of standardized test sets or evaluation methods make sentence segmentation a complex and demanding task.

A significant development in sentence segmentation is the introduction of self-supervised multilingual sentence segmentation methods. These methods are trained on unsegmented text, using newline characters to implicitly perform segmentation into paragraphs.[12] Such approaches are beneficial for low-resource languages and diverse corpora, as they do not rely on punctuation and require minimal sentence-segmented training data.

Accurate sentence segmentation is important for downstream tasks, like diarization.

Diarization

Speaker diarization, also known as speaker identification, is the process of partitioning an audio stream into homogeneous segments according to the speaker's identity. This task is particularly pertinent in situations involving multiple speakers, such as conference calls, interviews, or group discussions, where the goal is to determine "who spoke when."[4]

Several algorithms and techniques have been developed over the years for speaker diarization. The early approaches primarily relied on statistical methods for speaker change detection and clustering of speech segments. Generalized likelihood ratio and Bayesian information criterion were popular techniques used during the initial stages of diarization technology development.[4]

With the advent of deep learning, significant advancements have been made in speaker diarization. ML-based methods, such as extracting speaker embeddings using deep neural networks, have shown promising results. These methods leverage the powerful modeling capabilities of neural networks for speaker diarization, enabling easier training, enhanced performance, and robustness against speaker variability and acoustic conditions.[4]

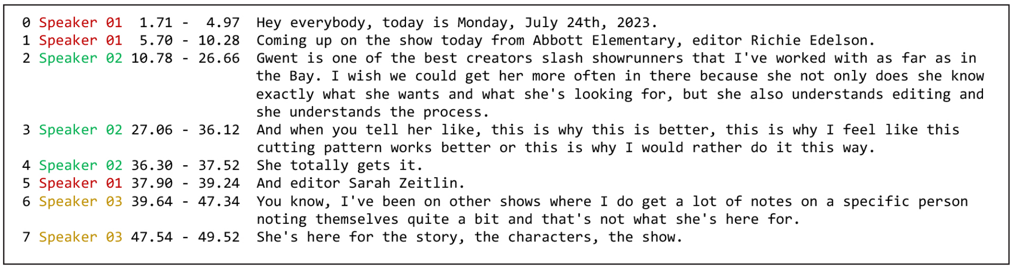
Clustering is a crucial component of speaker diarization, aimed at grouping speech segments based on speaker identity. Spectral clustering is a commonly used approach for this task that groups samples based on how similar they are to each other. This method involves several steps, including calculating an affinity matrix, creating a Laplacian matrix, eigen decomposition, re-normalization, speaker counting, and spectral embedding clustering[4] as shown in **Fig. 7**.

In the diagram to the left, you can see how the speaker embeddings from a conversation with three people are reduced to a 2D plot in clusters. Note that the cluster in the upper right is for samples of the same person speaking with a different tone.

The plot in the center shows a graph of the Silhouette Coefficient, a measure of cluster quality, with varying numbers of clusters (potential speaker counts.) The probable number of speakers in the recording is typically the one that maximizes the Silhouette Coefficient, which is 3 for this example. This indicates a scenario where audio segments are most similarly grouped within the same speaker and most distinct between different speakers.

The plot on the right shows the speaker assignments for the embeddings. The transcript with the identified speakers is shown in **Fig. 8**.

Note that this clustering method assumes equal speaker



```
0 Speaker 01  1.71 -  4.97  Hey everybody, today is Monday, July 24th, 2023.
1 Speaker 01  5.70 - 10.28  Coming up on the show today from Abbott Elementary, editor Richie Edelson.
2 Speaker 02 10.78 - 26.66  Gwent is one of the best creators slash showrunners that I've worked with as far as in
                            the Bay. I wish we could get her more often in there because she not only does she know
                            exactly what she wants and what she's looking for, but she also understands editing and
                            she understands the process.
3 Speaker 02 27.06 - 36.12  And when you tell her like, this is why this is better, this is why I feel like this
                            cutting pattern works better or this is why I would rather do it this way.
4 Speaker 02 36.30 - 37.52  She totally gets it.
5 Speaker 01 37.90 - 39.24  And editor Sarah Zeitlin.
6 Speaker 03 39.64 - 47.34  You know, I've been on other shows where I do get a lot of notes on a specific person
                            noting themselves quite a bit and that's not what she's here for.
7 Speaker 03 47.54 - 49.52  She's here for the story, the characters, the show.
```

Figure 8. Transcript with recognized speakers.

distribution and speaking time throughout the recording, which may not always hold true, necessitating the need for more advanced analysis methods.[13]

Despite significant progress, several challenges persist in speaker diarization. One of the primary challenges is handling overlapping speech, where two or more speakers are speaking simultaneously. Traditional diarization systems often struggle to identify and separate the speakers in such scenarios accurately.

Another challenge lies in dealing with the variability in speech characteristics across different speakers, including differences in accent, speaking rate, pitch, and volume. These variations can significantly impact the performance of diarization systems, making it difficult to accurately label and separate speech segments.

Post-processing

The output of ASR systems typically comes in the form of raw transcriptions. However, these transcripts often need to be converted to standardized formats to be useful in practical applications, especially in media production. Post-processing is the step where the raw transcript transforms to suit specific requirements. One of the primary tasks in post-processing is converting the ASR-generated transcript into a standardized closed caption or subtitle format. This ensures compatibility with different media platforms, broadcasters, and playback devices.[14]

The choice of closed caption format largely depends on the target platform and the intended use of the media content. Here's a brief overview of the primary considerations for each popular captioning formats.

When compliance with FCC broadcaster regulations is a concern, or if compatibility with CEA-608 (line 21)[15] and CEA-708[16] caption capabilities is needed, the SMPTE-TT[17] format is ideal. The EBU-STL[18] format is suited for PAL broadcasts in Europe. Given its frame-based nature, ensuring the timings match the video's frame rate is vital. EBU-TT[19] is slowly replacing this format but it remains relevant due to its versatility.

TTML[20] is a flexible choice for online video content, especially if the media will be uploaded to online platforms. Its time-based nature makes it versatile across different video frame rates.

Here are the steps in post-processing the transcripts for closed captioning files:

1. **Formatting & Styling:** Depending on the chosen format, specific styles, fonts, and positioning details can be applied. For instance, SCC subtitles have a character limit per line that must be adhered to, while EBU-STL allows for varied fonts and styling.
2. **Conversion to Desired Format:** The styled and segmented transcript is converted to the desired closed caption format using specialized software or tools. This step ensures the transcript adheres to the technical specifications of the chosen format.
3. **Quality Assurance:** The generated caption file is tested for synchronization, accuracy, and readability. Any discrepancies or errors are corrected at this stage.
4. **Export & Delivery:** The final closed caption file is exported and delivered in the required format, ready for integration with the video content.

Post-processing is essential in the ASR pipeline, especially when the end goal is media production. By converting raw transcripts into standardized closed caption formats, content creators can ensure their media is accessible, compliant, and ready for diverse audiences and platforms. However, note that caption generation faces challenges in depicting descriptive audio, including background noises and the tone of speech, especially when the speaker isn't visible. As media consumption grows and diversifies, ensuring content is appropriately captioned will remain a priority, making the post-processing step in ASR workflows even more crucial.

Using Machine Learning for ASR

In recent years, the introduction of new deep-learning models has significantly changed the NLP field.[21] Initial works in this field used Hidden Markov Models (HMMs),[22] Convolutional Neural Networks (CNNs),[23] and Recurrent Neural Networks (RNNs).[24] These newer works significantly enhanced the performance and accuracy using Transformer models,[25] which can analyze and understand the long-range dependencies within a given signal. Moreover, these new models, including OpenAI's Whisper,[1] Meta's MMS,[2] and Google's USM,[3] support more languages than older methods.

In the following sections, we describe transformers and then explain these new transformer-based models designed for ASR tasks.

What are Transformer Models?

Sequential data often contain dependencies, where understanding one part can be crucial to understanding another. Traditional neural network designs like HMMs and RNNs were developed to handle such sequential data. However, they often struggled with long-term dependencies, vital in language modeling tasks. Enter the Transformer architecture, which

1545-0279/24©2024SMPTE

effectively addresses this limitation. A unique feature of the Transformer is its separation into encoders and decoders. The encoders process the input sequence, while the decoders produce the output, which is especially vital in tasks like translation, where an input in one language is transformed into an output in another language,[25] as shown in **Fig. 9**.

At the heart of both the encoder and decoder components of the Transformer model is the attention mechanism.[26] Initially introduced for NLP tasks like translation, attention allows the model to selectively focus on specific portions of an input sequence, irrespective of the distance between elements. This selective focus is achieved by assigning different levels of importance or "weights" to different parts (or tokens) of the sequence. The more relevant a token is to the current token being processed, the higher its importance.

The Transformer further amplifies its power by simultaneously employing multiple such attention mechanisms, known as "attention heads." These heads allow the model to concentrate on various aspects of the input, enabling it to capture different nuances and relationships within the data. By leveraging multiple attention heads in its encoder and decoder structures, the Transformer becomes incredibly efficient and versatile in handling diverse NLP tasks.

State-of-the-art Models for ASR

Let's look at some of the most advanced models developed by leading tech companies, including OpenAI's Whisper, Google's USM, and Meta's MMS.

OpenAI's Whisper

Introduced by OpenAI in 2022, Whisper supports transcription and translation for 97 languages. It was trained on 680K hours of labeled audio data sourced from the web. This model can detect languages and transcribe or translate audio, providing word timings, capitalization, and punctuation for most supported languages. It is available in five sizes: tiny, base, small, medium, and large.

Google's USM

Google USM, also known as Chirp,[27] was launched in 2023. It underwent pre-training on an expansive collection of 12 million hours of unique YouTube audio spanning 300 languages. Subsequently, it was fine-tuned with a labeled dataset of 90K hours to master ASR in up to 100 languages. The model is adept at performing speech transcription and translation. Additionally, USM offers auto punctuation and word timings for many of its supported languages.

Meta's MMS

Developed by Meta and released in 2023, the MMS model expanded the range of supported languages for speech recognition from 100 to over 1,000. Impressively, it can identify more than 4,000 languages. MMS offers three distinct pre-trained models: one for speech-to-text, another for text-to-speech, and a third for language identification. It is available in two model sizes: one with 300 million parameters and another with 1 billion parameters. It's worth noting that MMS does not provide punctuation and capitalization.

Table 1 provides a summary of the features of these three models.

In addition to open-source models described in academic papers, several companies offer speech-to-text services. Notable platforms include Microsoft Azure's Speech to Text,[28] Amazon Web Services' Transcribe,[29] and Google Cloud's Speech-to-Text.[30] **Table 2** provides a comparative overview of their features.

Metrics

Word Error Rate (WER) and Character Error Rate (CER) are accuracy metrics used to evaluate the performance of ASR

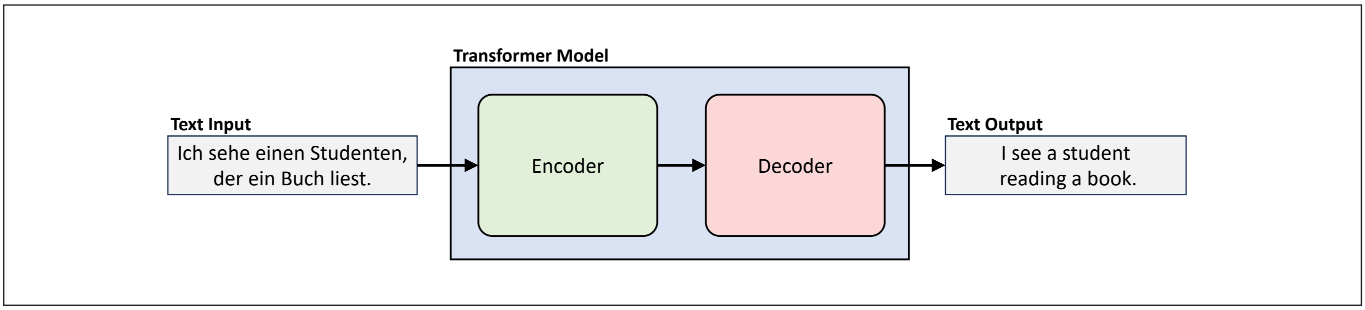

Transformer Model
Text Input
Ich sehe einen Studenten, der ein Buch liest.
Encoder
Decoder
Text Output
I see a student reading a book.

Figure 9. Components in a transformer model.

Table 1. Comparison of ASR models.

	Transcription	Translation	Supported Languages	Language Identification	Word Timestamps	Auto Punctuation	Open-Source
Whisper	✓	✓*	97	✓	✓	✓	✓
USM	✓	✓	100	✗	✓	✓	✗
MMS	✓	✗	>1000	✗	✓	✗	✓

*Note that the Whisper model only translates from other languages to English.

techniques.[31,32] Both metrics consider substitutions, deletions, and insertions in their calculations, comparing the ASR output against a reference,[33] as shown in **Equations 1 and 2**.

$$WER = \frac{\text{Number of word errors (substitutions, deletions, and insertions)}}{\text{Number of words in the reference}}$$

Equation 1. Word Error Rate Calculation

$$CER = \frac{\text{Number of character errors (substitutions, deletions, and insertions)}}{\text{Number of characters in the reference}}$$

Equation 2. Character Error Rate Calculation

WER assesses the accuracy at the word level, with a lower rate indicating superior ASR performance. Typically, a WER of 5-10% is deemed good, while up to 20% is acceptable. CER operates similarly but evaluates accuracy at the character level. A CER of 1-5% is good and up to 10% is acceptable.

To illustrate, let's consider a classic sentence, "The quick brown fox jumps over the lazy dog." If a system were to transcribe or translate it with minor errors such as "The qoick brwn fox jummps over the lazy dog," we find that the WER is 33.3% and the CER is 6.82%. This means a third of the words and just under 7% of the characters are wrong and would need fixing to match the original sentence.

Experiments

In this section, we present a summary of our experimental results testing transformer models and ASR APIs. The tests were performed using the Common Voice dataset[34] and included English and 16 non-English languages. We also used the Fleurs dataset[35] for Korean and Telugu since the Common Voice dataset didn't cover these languages. Sample sizes were chosen based on the total number of speakers per language,[36] resulting in 4,000 English and 9,000 multilingual samples.

To test the open-source models, MMS and USM, we conducted our experiments in a controlled environment leveraging a Python framework on a Windows operating system equipped with an NVIDIA RTX A6000 GPU. The models and datasets were publicly available and sourced from the Huggingface platform. We stripped any punctuation, special characters, and redundant spaces from the ground truth texts and the model outputs to make the metric calculations more robust. Then, we calculated the evaluation metrics, WER and CER, using the TorchMetrics library.

For the evaluation of the online services, AWS, Azure, and GCP, we used Python code to process identical audio files for transcription. Given audio format requirements, we converted the audio files from MP3 to WAV using FFmpeg.

We calculated the transcription speed for all tests by dividing the audio duration by the processing time, reporting the average speed across all samples. Setting up a proper code structure and conducting these tests across all services and models required one month.

The performance of the OpenAI Whisper large-v2 model, MMS-1B-all model, and ASR services are displayed in **Table 3** with the best results for local models and online services are in bold.

The results show that MMS outperforms Whisper in non-English languages, while Whisper shows superior WER/CER metrics for English. Note that we specified the language for all these tests. The Whisper model is unique because it can run in a mode that auto-detects the language. However, based on our experiments, the WER increases by about 3.6% using this mode. Also, MMS shows better speed compared to Whisper. We couldn't directly test the USM model as it's not publicly available. We used the results provided in the MMS

Table 2. Comparison of ASR services.

	Supported Languages	Language Identification	Word Timestamps	Auto punctuation	Diarization
AWS	~27	✓	97	✓	✓
Microsoft Azure	>100	✓	100	✓	✓
Google Cloud	>70	✓	>1000	✓	✓

*Note that the Whisper model only translates from other languages to English.

Table 3. Experimental results of MMS-1B-all, OpenAI Whisper large-v2, USM, and ASR APIs.

		Metric				
		WER		CER		Transcription Speed (Faster Than Realtime)
System Type	ASR System	English	Multilingual	English	Multilingual	
Local Model	MMS	20.10	26.47	8.29	8.98	54.74x
	Whisper	8.86	26.69	7.11	17.09	9.90x
	USM	NA	NA	4.60	7.14	NA
Online Service	AWS	7.75	22.21	9.00	14.45	0.23x
	Azure	10.32	16.27	6.11	6.20	4.73x
	GCP	29.58	25.30	19.89	11.54	4.18x

paper, which includes CER values for each language. Our analysis, based on a weighted average across these languages, indicates that USM has superior CER metrics for both English and multilingual content.

Our experiments also found that AWS offers the most accurate WER metrics for English content, but it is the slowest processing speed. Microsoft Azure, on the other hand, shows the best WER results for Multilingual and the best CER results for English and multilingual audio files. However, it requires silence removal using VAD. Azure also has a limit on the number of transcription tasks per minute.

Conclusion

The functionality provided by ASR is core to media production. Automatic captioning, searching media, and script production and alignment all benefit from ASR. We have presented an overview of the various steps in ML-based ASR, including Pre-processing, VAD, Language Detection, Speech-to-text, Sentence Segmentation, Diarization, and Post-processing. The role of the transformer is highlighted as a recent innovation that allows for accurately identifying long-term dependencies.

Several models were compared based on WER and CER, and results were presented for each. OpenAI's Whisper, Google's USM, and Meta's MMS were evaluated, in addition to the ASR services provided by AWS, Azure, and Google Cloud.

Acknowledgements

We would like to thank Jennifer Lim and Shailendra Mathur for reviewing our paper and providing valuable feedback.

References

1. A. Radford, et al.,"Robust Speech Recognition via Large-Scale Weak Supervision," ArXiv, abs/2212.04356, 2022.
2. V. Pratap, et al., "M. Scaling Speech Technology to 1,000+ Languages," ArXiv, abs/2305.13516, 2023.
3. Y. Zhang, et al., "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages," ArXiv, abs/2303.01037, 2023.
4. Park, Tae Jin, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan, "A Review of Speaker Diarization: Recent Advances with Deep Learning," *Comp. Speech & Lang. 72* (2022): 101317, 2022
5. S. S. Stevens, J. Volkmann, and E. B. Newman. "A Scale for the Measurement of the Psychological Magnitude Pitch," *J. of the Acoust. Soc. of America 8(3)*: 185-190, (1937.
6. E. Sejdić, I. Djurović I., J. Jiang (2009). "Time-frequency Feature Representation Using Energy Concentration: An Overview of Recent Advances," *Digital Signal Processing, 19(1)*: 153-183. doi:10.1016/j.dsp.2007.12.004.
7. Google. WebRTC, 2011. Accessed Aug. 2023. [Online]. Available: https://webrtc.org/
8. S. Team, "Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," 2021.
9. M. Bhatia, et al., "VoIP: An In-Depth Analysis - Voice Activity Detection," Cisco, Oct. 2006
10. M. Heck, Automatic Language Identification for Natural Speech Processing Systems, *The Department of Informatics Institute of Anthropomatics (IFA) Interactive Systems Laboratories (ISL)*, 2011.
11. R. Wicks and M Post, "A unified approach to sentence segmentation of punctuated text in many languages," *Proc. 59th Ann. Meeting of the Assoc. for Comput. Ling.*, pp. 3995-4007, 2021.
12. B. Minixhofer, et al., "Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation," arXiv preprint arXiv:2305.18893, 2023.
13. Park, Tae Jin, et al. "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, 27:381-385, 2019.
14. Graham Jones, "Implementing closed captioning for DTV," *Proc. Broadcast Engineering Conference*, National Association of Broadcasters (NAB), p. 8, 2004.
15. Society of Cable Telecommunications Engineers (SCTE), SCTE 21 2012 – "Standard for Carriage of VBI Data in Cable Digital Ttransport Streams," SCTE 21: 13. 2012.
16. Consumer Technology Association (CTA)/American National Stndards Institute (ANSI), "Digital Television (DTV) Closed Captioning" (ANSI/CTA-708-E S-2023) 2013.
17. SMPTE, ST 2052-1:2010, "Timed Text Format."
18. European Broadcasting Union (EBU), "Specification of the EBU Subtitling Data Exchange Format," PDF. February 1991. Retrieved 10 Mar. 2013.
19. European Broadcasting Union (EBU), Tech 3380, EBU-TT-D Subtitling Distribution Format Version 1.0.
20. "Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP)," 2018.
21. A. Mehrish, N. Majumder, R. Bhardwaj, R. Mihalcea, and S. Poria, "A Review of Deep Learning Techniques for Speech Processing," *Information Fusion*, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2305.00359
22. G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. "Multilingual Acoustic Models Using Distributed Deep Neural Networks," *Proc. of ICASSP*, 2013.
23. Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*:1533-1545, 2014.
24. Alex Graves, Abdel-Rahman Mohamed, Gregory Hinton, "Speech Recognition with Deep Recurrent Neural Networks," ICASSP, Proc. *IEEE Intern. Conf. on Acoustics, Speech and SIgnal Process*. 38. 10.1109/ICASSP.2013.6638947, 2013.
25. A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need, NIPS," 2017.
26. D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," CoRR, abs/1409.0473, 2014.
27. Chirp: University Speech Model. [Online]. Available: https://cloud.google.com/speech-to-text/v2/docs/chirp-model
28. Azure: Speech to text. [Online]. Available: https://azure.microsoft.com/en-us/products/ai-services/speech-to-text#features
29. Amazon Transcribe Features. [Online]. Available: https://aws.amazon.com/transcribe/features/?nc=sn&loc=2
30. Google Cloud's new visual interface for Speech-to-Text API | Google Cloud Blog. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/google-clouds-new-visual-interface-for-speech-to-text-api
31. A.C. Morris, V. Maier, and P.D. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," Interspeech, 2004
32. Hugging Face. [Online]. Available: https://huggingface.co/learn/audio-course/en/chapter5/evaluation#character-error-rate
33. Learn. Azure. Test Accuracy of a Custom Speech Model. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio#resolve-errors-and-improve-wer
34. R. Ardila, et al. "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912*.06670, 2019.
35. A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, C., and A. Bapna, "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech," IEEE Spoken Language Technology Workshop (SLT), pp. 798-805, 2022.
36. List of languages by total number of speakers. [Online]. Available: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

About the Authors

Randy Fayan, Sr. Director of Engineering at Avid, has worked in the media industry for more than 25 years. He has a long history of leading teams to bring innovative features to market. His current focus is on adapting AI/ML solutions tailored to the needs of media professionals.

Rob Gonsalves joined Avid as their 15th employee in 1989. He helped develop the industry's preeminent nonlinear editing system Avid Media Composer, specializing in programming video effects, color correction. His research focuses on AI for media production. Gonsalves holds over 50 patents in the media and entertainment industry.

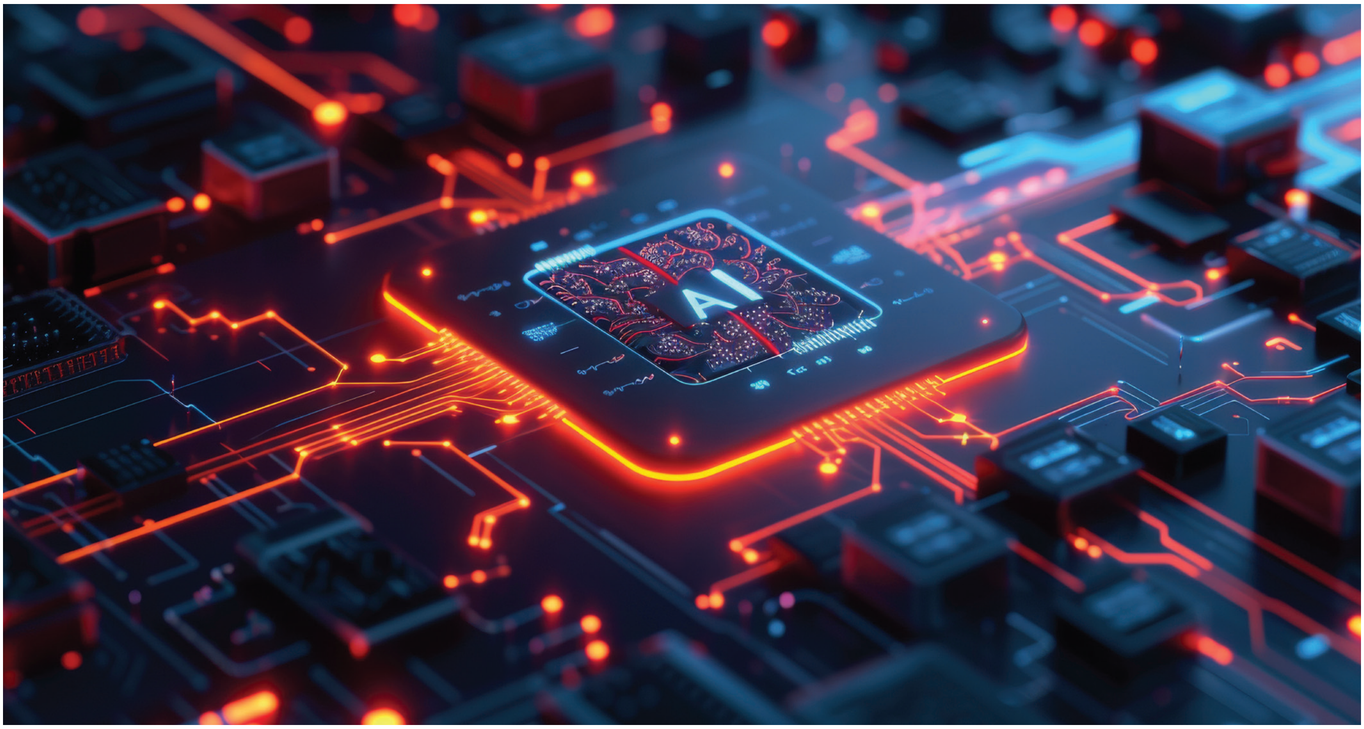
Zahra Montajabi is an AI researcher intern at Avid Technology. Currently, she is AI/ML software developer in Montreal, Canada. Her research interests are computer vision, image processing, and machine learning.

Presented at the 2023 Media Technology Summit. Hollywood, CA, October 16-19, 2023.
Copyright © 2024 by SMPTE.

DOI: 10.5594/JMI.2024/IPYX8877
Date of publication: 1 April 2024

APPLICATION/PRACTICES

KEYWORDS ARTIFICIAL INTELLIGENCE // MACHINE LEARNING // AUTOMATIC SPEECH RECOGNITION



AI-Powered Editorial Systems and Organizational Changes

D. Arets, M. Brugman, and J. de Cooker

Although editorial systems are the heart and backbone of the journalistic process, they are "underexposed" within innovation processes and journalistic research. This is remarkable, because technological developments, especially the rapidly evolving field of Artificial Intelligence (AI), promises numerous opportunities for revitalizing editorial systems.

1545-0279/24©2024SMPTE

Abstract
Journalists work with editorial systems daily, using them to process their articles, communicate with colleagues, and produce and archive content. Yet even though editorial systems are the heart and backbone of the journalistic process, they are "underexposed" within innovation processes and journalistic research. This is remarkable, because technological developments, especially the rapidly evolving field of artificial intelligence (AI), promises numerous opportunities for revitalizing editorial systems, as well as new (hybrid) ways of working, and needs to be considered in terms of redesigning existing editorial systems and workflows. In the research undertaken in this initiative, a mixed group of software engineers, AI experts, and journalism and design researchers, collectively referred to as the "The Editorial Portal" was assembled at the Netherland's Fontys University of Applied Sciences. This group was tasked with investigating opportunities for a future-orientated editorial system in which both organizational and technological transformations were considered. The study was undertaken with the understanding that design methods, including context mapping, were suitable for identifying the relationship between editorial system, corporate culture, and future developments. As it was felt that working professionals should be involved, journalists from four Dutch regional newsrooms were also consulted in this study. They identified a need for a more intelligent system that encourages collaborative and creative working methods.

Editorial systems are underexposed in journalistic research, which seems strange as such systems form the basis of the journalistic process. The sparse studies available stress that editorial systems are the backbone of the journalistic process. According to Holmberg,[1] editorial systems serve multiple essential purposes, including equipping editors with toolkits for quickly selecting and sorting information, minimizing parallel work, facilitating efficient information flow between staff members, controlling production workflow, and managing scheduling and resources. Further, editorial systems are crucial for newsroom collaboration as they facilitate work processes and communication between journalists, final editors, and production according to Marjoribanks.[2,3]

The crucial role of editorial systems came to light during the Covid-19 pandemic when journalists were forced into hybrid work modes. Recent research by the Reuters news agency indicates that this hybrid work heavily impacted the newsroom culture, with those working in the newsroom and those working from home, participating unequally in the newsgathering and publishing processes. This inequity resulted in organizational and technical challenges, among those with editorial systems, as current systems are not well-equipped to facilitate hybrid newsroom operations.[4] Many news operations in the Netherlands adopted Microsoft Teams to enable working from home and video conferencing.

In addition, artificial intelligence's rapid advancement is challenging journalism in all its facets, including the journalistic editorial system. AI-driven tools for transcribing, translating or image editing, have been introduced in the newsroom. However, the current editorial systems in Dutch regional news outlets are not yet really equipped for working with AI.

The need for new, more hybrid operating systems, coupled with the rapid evolutions around AI, prompted Dutch regional news outlet Omroep Flevoland, in collaboration with regional outlets Omroep West, RTV Utrecht, Omroep Rijnmond, and Omroep Zeeland, to commission Fontys University of Applied Sciences to investigate future-orientated news systems.

During this one-year research project, Fontys researchers in the department of Professorship Journalism and Innovation teamed up with their peers in the Design and Interaction department to form a team consisting of two journalism researchers, one software engineer, two user experience (UX) designers, and one innovation researcher. Four journalism students, 12 software engineering students, and 10 media design students also took part in the research project.

The primary objective of this project was to ascertain the organizational, technical, and functional requirements that must be incorporated in a future journalistic editorial system to adequately and responsibly respond to hybrid and more AI-orientated news practices.

In order to determine this, four research groups were created within the team:

The first, "Group 1," consisted of researchers from the Professorship of Journalism and Innovation, and focused on mapping existing journalists' workflows with the edito-

1545-0279/24©2024SMPTE

rial system and determining specific requirements for a future-orientated system.

The second, "Group 2," was comprised of software engineering students and lecturer-researchers from the Professorship of Design and Interaction, and was tasked with investigating the system design that would be necessary for an AI-driven editorial system.

The third, "Group 3," was made up of Design and Interaction students and a lecturer-researcher from that department, and investigated the UX requirements of such an AI-driven system.

The last team, "Group 4," was formed from principal investigators of all teams and project managers from the participating news outlets. This group investigated editorial system journalistic organizational developments.

The groups worked in parallel and exchanged results and insights through six-weekly meetings. In this way, research on mapping the journalistic process by research Group 1 was supported by the UX research of Group 3 and vice-versa. The process of mapping the software structure of research Group 2 fueled the qualitative interviews with journalists by researchers of Group 1, and challenged Group 4 to consider future-orientated organizational structures.

Although all groups used their own research methods (ranging from more ethnographic and qualitative research for Group 1, AI engineering in Group 2, and UX design research in Group 3), all four groups followed a research trough design process, where prototyping and development of tools went together with reflection.[5]

The combined insights from all four groups have been processed and are presented here in three phases:

Phase 1: Mapping Ways Of Working With Editorial Systems

Regional broadcasters in adjacent regions in the Netherlands cooperate, sharing data and exchanging knowledge for organizational development. In addition, there is cooperation with the country's National Broadcasting Foundation (NOS). However, where editorial systems are concerned, there is little such cooperation. This is understandable as there are four different editorial systems in use by regional broadcasters, with some using the Dalet system, while others use Nimbus or Nis, and still others work with the INOS system that was developed and is used by the NOS.

To properly understand and interpret how journalists interact with their editorial systems, a combination of "desktop walkthroughs"[6] and semi-structured interviews was used in the first phase of the research. A "desktop walkthrough" is a valuable tool, as it provides an opportunity for discussions about everyday actions that often occur intuitively or subconsciously. In this way, even seemingly trivial elements or intuitive operations that often remain undiscussed in an interview setting can be addressed.

Twenty-two newsroom professionals were observed and interviewed. These consisted of online editors, anchors, archivists, producers, and designers from four regional news outlets. These journalists were asked to describe their daily activities such as starting up a newsroom system, opening tabs into the editorial portal, communicating with colleagues, editing an item for social media and how this item would be processed for a TV, radio and online distribution, and finally the archiving of an item. This entire process was recorded via smartphones (**Fig. 1**).

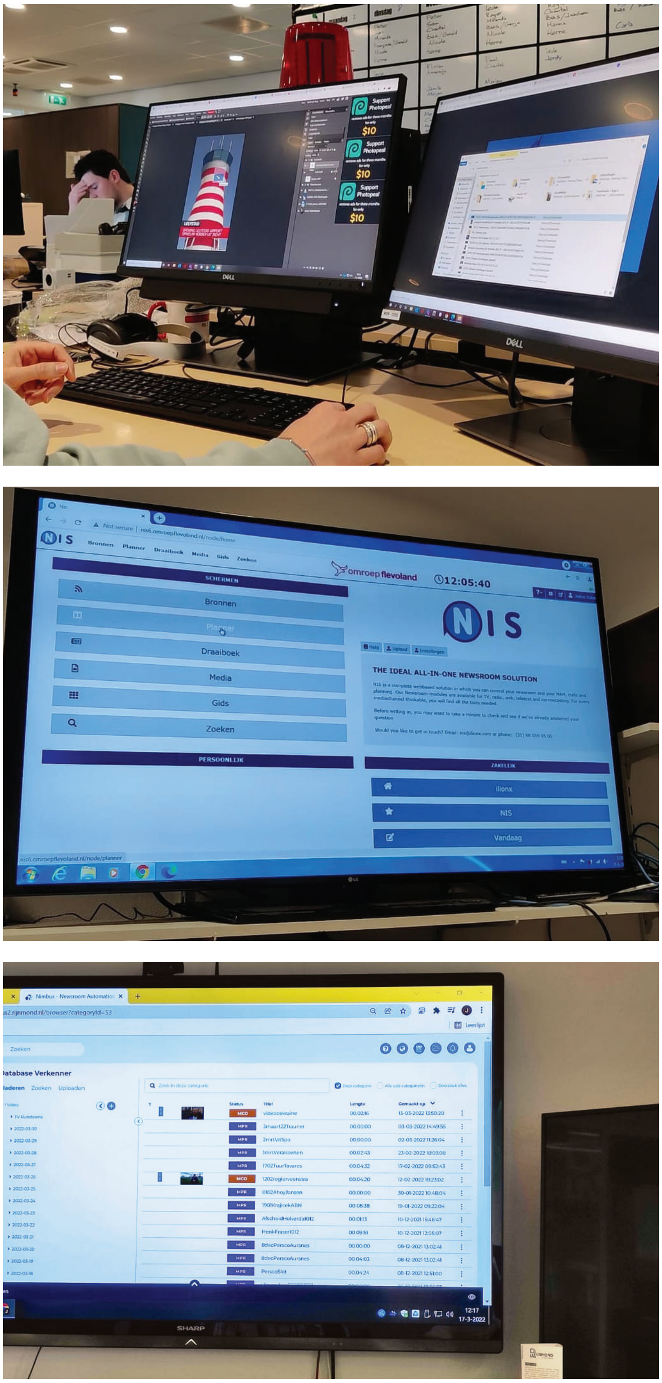


Figure 1. Desktop walkthrough at four Dutch regional news outlets.

Subsequently, the journalists were interviewed using a topic list. They were asked to discuss how the editorial system facilitates current work processes, its merits, and possible frustrations that it creates, as well as to describe annoyances associated with the existing portal, and their desires for a future strategy. Excerpts from these recordings were discussed during the interview in order to clarify operations

1545-0279/24©2024SMPTE

or “workarounds” that may have been created. In the discussions, emphasis was placed on how routines changed during the pandemic when many newsrooms began using Microsoft Teams to facilitate remote working.

Insights Phase #1

Opportunities for better streamlining editorial systems

With the exception of two of the journalists consulted, who described their systems as “fine and pleasant,” most interviewees expressed indifference to their editorial system and four of the respondents were reluctant to discuss their editorial systems, stating that they had no say in such matters, or it would be better to speak with their technical colleagues.

These reactions correspond with those reported in research conducted by Brautovic,[7] which indicates resistance to editorial systems stems from users (journalists) having minimal or no involvement in their development and implementation, and producing a feeling of “no ownership.” Brautovic additionally pointed out that most systems are dominantly tech-orientated, and as a result, many applications are not tailored for journalistic practices.

The latter aspect also became evident from analysis of the “desktop walkthrough.” It was found that the vast majority of journalists performed operations outside of the editorial system. For instance, all consulted journalists had at least three extra tabs open in addition to the editorial system itself. These might involve an online search engine or a tab for police information channel or a communication tab (primarily social media). Moreover, additional software programs were active in the users’ browsers; typically Excel for data processing, Photoshop for image editing, Adobe for video editing, and Microsoft Teams for video conferencing.

All 22 journalists consulted indicated a desire for the editorial system to better integrate the various software and distribution channels involved.

When discussing the design and interface of the editorial system, at least six journalists mentioned that they prefer working with a “clean sheet,” as their current system included numerous functionalities that sometimes caused distraction and upset. One responded that the system they used was developed from a technological viewpoint, where functions were designed without giving consideration to the end user. The preference was for a “simple and clean” system.

AI-empowered newsrooms

The number of computer mouse “clicks” necessary to achieve a task was among the most frequently cited frustrations during the interviews. A journalist must repeatedly “click” between the mandatory input fields to create a journalistic item. An online article has a headline, a subhead, a lead paragraph, an image, and an author, and all of these elements have a specific input field, with each requiring a separate “click” of the mouse. A particular challenge was in archiving. More than half of the consulted journalists mentioned the archival filing of an article requires an excessive amount of time and creates frustrations. The logging of a report with the required metadata was also cited as requiring an excessive amount of time to perform.

At the same time, all interviewees indicated that their digital archives were difficult to search, which was attributed to poor maintenance of the archive, with those editors using it neglecting to add necessary metadata to their filings. One of the consulted archivists mentioned that a lot of work and frustration could be eliminated if metadata could be added automatically.

In addition to a “smart archive,” opportunities were seen for application of AI in other parts of the editorial system. For example, 14 out of the 22 interviewees thought it could be used to create a better system for publishing news to various channels, with automatic tailoring of the content and images for a specific medium.

Communication & Collaboration

Research by Marjoribanks[8,9] indicates that editorial systems impact working relations. An editorial system interacts with working methods and the routines of people, and defines the conditions for interaction. Robinson found that virtual environments correlate with physical ones.[10] Online spaces operated as extensions of the physical newsroom, and relationships that develop in such “virtual platforms” translate into the physical newsroom. However current editorial systems support the relationship between the physical and online domains. At least that was evident during the pandemic when working relationships were disrupted with those working remotely lacking alignment with the physical workspace.[4]

For more than half of the consulted journalists, current editorial systems do not adequately support communication, collaboration, or planning. In the observed newsrooms, only one primarily used the editorial system for communication. One newsroom employed a separate Trello system for communication and team planning, and one newsroom communicated mainly via Microsoft Teams. All journalists consulted indicated that in addition to the established newsroom channels, they also communicated with colleagues via social media.

Although some interviewees acknowledged that the current style of working causes stress and irritation, they believed that it was inevitable and merely “a sign of the times.”

One respondent cited Microsoft Teams as a way to streamline communication better, observing that the calendar function and one’s communication were close together, and that it was possible to see if a colleague already had an appointment.

Another journalist mentioned that working with Microsoft Teams during the pandemic introduced the concept of multiple persons working together on a file, noting that such methodology is a prerequisite in supporting collaboration.

Phase #2: Mapping Future Editorial Systems

In all four of the regional news outlets investigated, experiments were being conducted involving the use of AI in the story discovery, research, and verification processes. However, such AI tools and applications have not yet been integrated in the editorial portals. As part of the study, journalists were queried about how they envisioned themselves in working with an AI-driven editorial system and what functionalities they thought such a system have, and also, how it might be useful in establishing new journalistic ways of working and practices. In investigating this, context mapping workshops were organized at the four news outlets.[11]

Context mapping captures contextual information of users interacting with the system. This form of generative research helps elicit emotional responses from the participants, reveals tacit knowledge, and exposes latent needs. Also, in contrast to more common research methods such as interviews or focus groups, which can reveal information about current and past experiences, the use of context mapping can provide insights into future desires as small evocative tasks that help to spark ideas about desired future environments.

One of the interactive workshop exercises involved the use of a card deck to help visualize the function and scope of AI-driven tools. Based on insights from the interviews and our AI inventory, eight such AI tools were designed.

Three of the tools are comparatively easy to integrate from both a technical and an organizational point of view. One involves the use of AI in archiving, automatically filing metadata associated with media productions, thus removing frustrating work and "clicking" from the journalist's hands. Another AI-driven tool could integrate various experts in connection with the news process, thus supporting the diversity of experts in the field. The third tool would be used for sentiment analyses of social media to quickly provide journalists with an overview of topical sentiments.

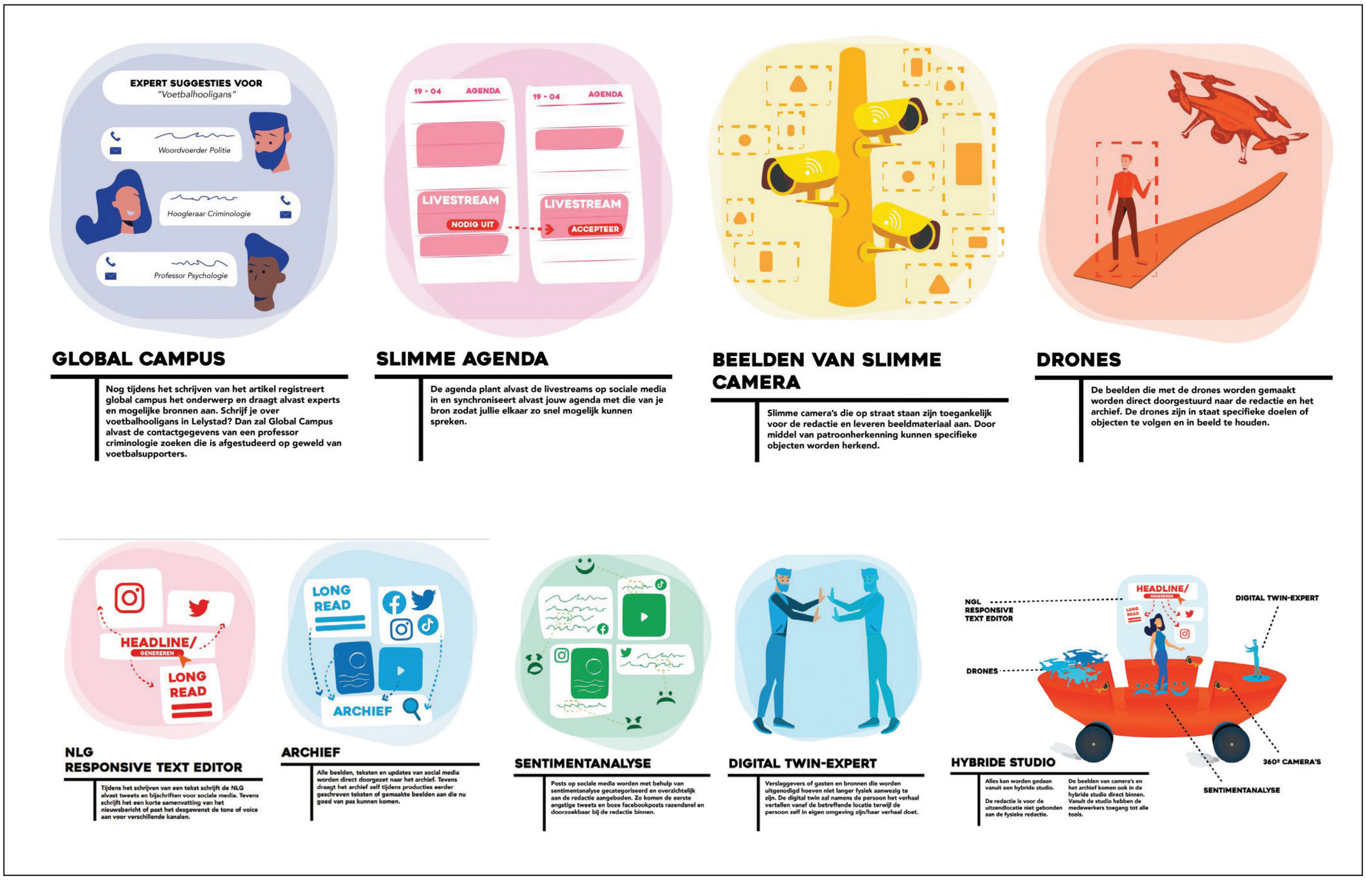
Another tool that would be easy to integrate from a technical point of view, but could be organizationally challenging, were a "smart planner" that would make appointments based on personal and team agreements and schedules, and thus eliminate unnecessary communication. This tool, however, would require all employees to share their calendars, including dates and times scheduled for personal meetings.

There was also a "responsive text editor" presented, which could write articles based on previous text, much like what is possible with Chat GPT. However, its implementation would require organizations to have a clear view as to how to responsibly operate with AI in text generation. As noted by Deuse and Beckett,[12] introducing AI in newsrooms calls for "algorithmic literacy" and "algorithmic transparency," as expressed by Diakopoulos and Koliska[13] along with "algorithmic accountability," as voiced by Arets, De Cooker, Wernaart.[14]

One tool that is easier to implement, yet technologically underdeveloped, involves a "digital twin" function that would allow an invited TV show guest to participate without actually being physically present.

Finally, two more future-oriented tools were presented: smart drones providing automatic recording of events and a hybrid news studio that can be situated anywhere.

In a one-and-a-half hour workshop, eight professionals from the regional news outlets (a mix of journalists, final editors, producers, and online editors) were asked to create two journalistic productions using the tools described. In addi-



GLOBAL CAMPUS

Nog tijdens het schrijven van het artikel registreert global campus het onderwerp en draagt alvast experts en mogelijke bronnen aan. Schrijf je over voetbalhooligans in Lelystad? Dan zal Global Campus alvast de contactgegevens van een professor criminologie zoeken die is afgestudeerd op geweld van voetbalsupporters.

SLIMME AGENDA

De agenda plant alvast de livestreams op sociale media in en synchroniseert alvast jouw agenda met die van je bron zodat jullie elkaar zo snel mogelijk kunnen spreken.

BEELDEN VAN SLIMME CAMERA

Slimme camera's die op straat staan zijn toegankelijk voor de redactie en leveren beeldmateriaal aan. Door middel van patroonherkenning kunnen specifieke objecten worden herkend.

DRONES

De beelden die met de drones worden gemaakt worden direct doorgestuurd naar de redactie en het archief. De drones zijn in staat specifieke doelen of objecten te volgen en in beeld te houden.

NLG
RESPONSIVE TEXT EDITOR

Tijdens het schrijven van een tekst schrijft de NLG alvast tweets en bijschriften voor sociale media. Tevens schrijft het een korte samenvatting van het nieuwsbericht of past het desgewenst de tone of voice aan voor verschillende kanalen.

ARCHIEF

Alle beelden, teksten en updates van social media worden direct doorgezet naar het archief. Tevens draagt het archief zelf tijdens producties eerder geschreven teksten of gemaakte beelden aan die nu goed van pas kunnen komen.

SENTIMENTANALYSE

Posts op sociale media worden met behulp van sentimentanalyse gecategoriseerd en overzichtelijk aan de redactie aangeboden. Zo komen de eerste angstige tweets en boze facebookposts razendsnel en doorzoekbaar bij de redactie binnen.

DIGITAL TWIN-EXPERT

Verslaggevers of gasten en bronnen die worden uitgenodigd hoeven niet langer fysiek aanwezig te zijn. De digital twin zal namens de persoon het verhaal vertellen vanaf de betreffende locatie terwijl de persoon zelf in eigen omgeving zijn/haar verhaal doet.

HYBRIDE STUDIO

Alles kan worden gedaan vanuit een hybride studio.

De redactie is voor de uitzendlocatie niet gebonden aan de fysieke redactie.

De beelden van camera's en het archief komen ook in de hybride studio direct binnen. Vanuit de studio hebben de medewerkers toegang tot alle tools.

Figure 2. Tools used in the context mapping workshop, designed by Floor van der Wal.

1545-0279/24©2024SMPTE

tion, participants were paired in two groups of four persons, each with varying levels of expertise (editor, planner, producer, technical expert), and asked to draw a story scenario utilizing the tools offered.

The first group was tasked with preparation of a news story (described in scenario #1 below), and the second group was asked to prepare an investigative report (described in scenario #2 below).

> *#Scenario 1: A report assumes football supporters have been terrorizing the region for weeks and social media research has revealed that a riot will take place soon. The challenge is to turn information about that riot into a report.*
>
> *#Scenario 2: In a press conference announced, the Minister of Housing & Spatial Planning, Hugo de Jonge, will declare that the established targets for building houses will not be met. Further, due to atmospheric pollution laws and high construction costs, planned large-scale projects are being halted or postponed.*

For 40 minutes, journalists had to investigate, create, and produce a news item utilizing the tools. For this, they were given all tool cards, with each card describing the particular tool and how it could be used. Then, the teams were asked to draw a story, deciding whether to use the tools described, and which team members should be involved in the creation of the story. Two researchers took notes of the conversation in both groups during the sessions. Then, the teams presented their stories in a plenary session, followed by a joint reflection session.

Insights Phase #2

All groups viewed the intelligent archive and the "smart planner" as immediately deployable and meaningful for the future news process. It was observed that AI-driven text tools, such as quickly compiling an initial template message with a responsive text editor and sentiment analysis, were positively considered by almost all groups (six out of eight) with little or no resistance concerning use of such technology. One of those with negative feelings mentioned that his place of work was not ready for automated text writing, stating that the organization would first have to become more knowledgeable about AI and that a checks-and-balance system would have to be implemented before moving to automated text writing.

The other respondent who expressed negative feelings about the technology stated that he feared it could jeopardize the creative aspect of his profession.

Most of the journalists consulted approached the use of AI-driven tools in a positive and cooperative manner, viewing this technology as a welcome addition that could minimize or eliminate rote tasks. They also saw opportunities for employing AI tools to adapt text for distribution via social channels. However, they all agreed that they needed to update their knowledge concerning the responsible use of AI to avoid biases or mistakes.

The smart drone camera was omitted in almost every scenario, except for that developed by one group. Most journalists mentioned that this tool has no added value, as drones are not allowed in public spaces, and also expressed concerns regarding privacy issues. One group, however, stated that they would like to use smart drones, especially in connection with covering dangerous events, such as when the soccer fans get out of control. However, that group also indicated that legislation should be amended to cover such eventualities.

All groups used the sentiment analysis tool. The groups working on the short report used it to quickly find out what the mood at the game would be like and which potential respondents should be interviewed. In the longer report, two groups deployed the sentiment analysis tool to support the research process. The two groups that did not use this tool indicated in the reflection session that while it would certainly be a desirable tool for providing initial scanning of dominant emotions surrounding a story, they expressed doubt concerning the credibility of such tools. One editor was very critical in his views, stating that use of such a tool might create complacency in its users and that vigilance and human judgement were needed to ensure against building reports from only social media reports.

Phase #3: Mapping The Future Newsroom

Finally, an investigation was conducted to explore how journalists might organize future news processes and design a system that will facilitate it. Participants were asked to prepare conceptual sketches as suggested by Van der Lugt.[15] This approach was combined with insights extracted from the semi-structured interviews performed in Phase 1 and the context mapping performed in Phase 2.

Journalists were asked to prepare sketches of their current work space and place themselves within this space. If there was hybrid working, this could be indicated with a dotted line. The resulting drawings were then discussed in a plenary session. The evaluation of the drawings helped to identify the most important parts of the newsroom. It was quite noticeable that the technical department was missing from many of the drawings. (However, this absence could be attributed to the location of the technical department on another floor, or possibly in another building. In some cases, editors admitted to forgetting to draw the operation's technical department.)

Almost all of the drawings either showed a fragmented structure, with groupings of various editors (social media sports, and news) or a newsroom structured around media distribution (radio, television, and online). Most indicated the editor-in-chief being in a separate location.

Next, editors were asked to draw their concept of what a newsroom might look like in 2030, with discussions among the editors taking place following the creation of the drawings. It was striking that almost all editors envisioned a workspace with more interconnectedness. The drawings showed fewer "islands," with more of a central desk where the journalism workers could "plug in and out," either locally or online during the workday. Some illustrations also featured a centrally-located coffee bar where colleagues could meet and discuss editorial matters.

None of the drawings showed that the physical newsroom

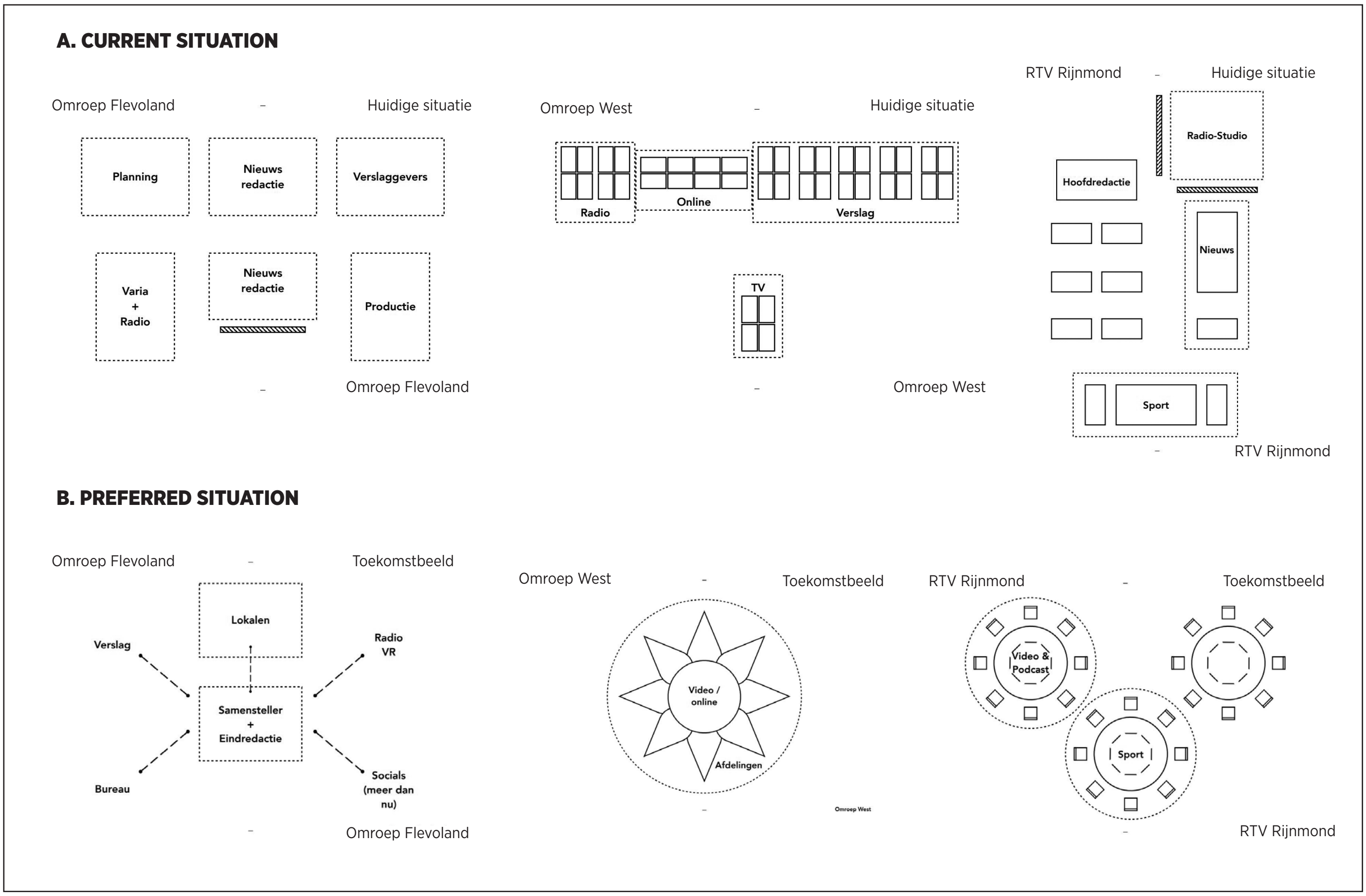


Figure 3. A indicates current situation of three newsrooms. B indcates preferred situation in 2030.

had become unnecessary. In fact, the physical space was considered especially important during the discussions. Editors stressed the importance of having a space where staff could interact during their daily activities. They suggested that future newsrooms should be designed in a way that further facilitates this interaction.

In discussing the drawings of the "future newsroom," the concept of interconnectedness seemed central, with many indicating that there should be more collaboration between various disciplines. Some participants' drawings depicted this quite literally, with the seating of multiple players (social media editor, journalist, and final editor) at one desk.

The journalists also indicated the need for greater collaboration on stories that would be distributed through various channels (TV, radio, and online), with editors from the various distribution channels sometimes working separately on the story. This aligns with research on the future of story-centric newsroom as described by Alba[16]

Finally, journalists involved in these "drawing" sessions have clearly stated that the editorial system developed must operate hand-in-hand with physical newsroom developments. This means that the system should also focus on mutual exchange, cooperation and the organization of centralized meetings.

Conclusion

The research described in this paper shows that journalists assume that future editorial systems will be AI-driven and that time-consuming tasks such as editing articles for various media channels, archiving media content, and organizational planning can be made easier with AI tools. However, there is hesitation in applying AI in the creation of textual content and in research and verification. Opportunities for AI tools exist in various areas but integrating them into an editorial system requires more knowledge and better control and monitoring of its operations.

Further, journalists consider it of utmost importance that the look and feel of a future system should be in line with the journalistic process without an overabundance of functionalities. The editorial system must also respond to future organizational developments in which journalists will increasingly collaborate and where multidimensional, rather than linear, storytelling is paramount. As the future newsroom will continue to work in a hybrid fashion — something that has become common practice since the pandemic — the future editorial system should facilitate this. This does not alter the fact that both the physical and digital news space, and thus the editorial system, will have to actively engage in collaborations and any future system must facilitate this. As

1545-0279/24©2024SMPTE

noted, editorial systems are underexposed in research, and it is the hope of the authors that this study will contribute to further investigate this area from a multidisciplinary point of view.

References

1. S. A. Holmberg, "Editorial systems for multiple channel publishing," Masters Thesis. Stockholm: The Royal Institute of Technology, 2002.
2. T. K. Marjoribanks, "News Corporation, Technology and the Workplace: Global Strategies, Local Change. Cambridge, Cambridge University Press, 2000.
3. T. Marjoribanks, "Strategizing Technological Innovation: The Case of News Corporation." In: Cottle S. (ed). Media Organization and Production, 59-75, Sage, London, U.K., 2003.
4. F. Cherubini, N. Newman, and R. Nielsen, "Changing Newsrooms 2021: Hybrid Working and Improving Diversity Remain Twin Challenges for Publishers," 2021.
5. A. Tashakkori and J. W. Creswell, "The New Era of Mixed Methods," *Journal of Mixed Methods Research*, 1(1): 3-7, 2007.
6. V. Auricchio, A. De Rosa, and M. Göransdotter, "Experiential Ways of Mapping: Revisiting the Desktop Walkthrough," Design International Series, 2022.
7. M. Brautović, "Usage of Newsroom Computer Systems as Indicator of Media Organization and Production Trends: Speed, Control and Centralization," *Medijska istraživanja: znanstveno-stručni časopis za novinarstvo i medije, 15*(1): 27-42, 2009.
8. T. K. Marjoribanks, "News Corporation, Technology and the Workplace: Global Strategies, Local Change. Cambridge: Cambridge University Press, 2000.
9. T. Marjoribanks, "Strategizing Technological Innovation: The Case of News Corporation." In: Cottle S. (ed). Media Organization and Production, 59-75, Sage, London, U.K., 2003.
10. S. Robinson, "Convergence Crises: News Work and News Space in the Digitally Transforming Newsroom. *Journal of Communication*, 61(6): 1122-1141, 2011.
11. F. S. Visser, P. J. Stappers, R. Van der Lugt, and E.B. Sanders, "Contextmapping: Experiences from Practice. CoDesign," 1(2): 119-149, 2005.
12. M. Deuze, & C. Beckett, "Imagination, Algorithms and News: Developing AI Literacy for Journalism. *Digital Journalism*, 10(10): 1913-1918, 2022.
13. N. Diakopoulos, and M. Koliska, "Algorithmic Transparency in the News Media," *Digital Journalism*, 5(7): 809-828, 2017.
14. D. Arets, J. De Cooker, "AI hoort in de beroepscode, NRC Handelsblad," Apr. 2023.
15. R. Van der Lugt, "How sketching can affect the idea generation process in design group meetings," *Design Studies*, 26(2): 101-12, 2005.
16. R. Alba, "The Challenges of Adapting News Production to the New Reality," presented at the SMPTE 2020 *Annual Tech. Conf. and Exhibit.* (Online) Nov. 2020.

About the Authors

Danielle Arets is Professor Designing Journalism at Fontys University of Applied Sciences / School of Journalism. The research group collaborates closely with media outlets to co-create the future of journalism.

Jessy de Cooker is a lecturer and researcher at Fontys Journalism. He focuses on AI in journalism and explores how new technologies can be deployed in line with journalistic values.

Marius Brugman is a researcher and lecturer at Fontys Journalism, where he looks at new technologies deployed in journalism. He questions and investigates the (im)possibilities that AI has to offer journalism.

Presented at IBC 2023, Amsterdam, The Netherlands, 15-18 September 2023. This paper is published here by kind permission of the IBC. Copyright © IBC

DOI: 10.5594/JMI.2024/RSMP6989
Date of publication: 1 April 2024

Join the Board of Editors
Volunteer to help shape and maintain the Journal's high editorial quality.
MI
MOTION IMAGING JOURNAL
SMPTE

1545-0279/24©2024SMPTE

KEYWORDS SMPTE ST 2110 // NMOS // SDN // PTP // SDP // MCS

PAC-12 Networks Builds a Brand-New Studio and Broadcast Center:

Utilizing State-of-the-Art SMPTE ST 2110 and Software-Defined Networking Solution

By Nik Kumar and Hieu Ho

Abstract

This paper describes the deployment of SMPTE ST 2110 suite of standards and networking technologies in constructing a brand-new technology center for PAC-12 Networks in San Ramon, CA. The project timeline spanned June 2022-September 2023. We discuss the underlying core technologies utilized, custom workflows developed, key challenges faced, and lessons learned.

PAC-12 Networks is a college sports digital cable and satellite television network owned by the PAC-12 Conference. PAC-12 Networks recently relocated to a brand-new facility in San Ramon, CA. They partnered with Advanced Systems Group, LLC (ASG)—a leading provider of solutions, systems integration, and services for media production and post-production headquartered in Emeryville, CA—to engineer, design and integrate a state-of-the-art technology center to support their on-air operations which, includes remote events production, studio production, broadcast ingest and transmission services, and replay and command center game officiation. PAC-12 Networks produces about 850 live sports events across over 100 venues during a season, mainly utilizing Remote Integration Model (REMI) production models (also referred to as Multicam in PAC-12 workflows) and in-facility production control rooms. Previously, PAC-12 Networks operated out of their old facility in San Francisco, CA, from 2012-2023.

At a high level, the new facility houses a main studio, five production control rooms, and five audio control rooms to support and produce concurrent sporting events. It also includes four hybrid software-defined production control rooms, a multipurpose facility that can be designated to serve in several aspects of live production (including bug and graphics operators, replay operators, ticker ops. etc.), and a technical operations center referred to as BITS. The core underlying technologies and equipment supporting these operational workspaces and workflows reside in a central equipment room housing 55 racks.

PAC-12 Networks utilizes leading-edge SMPTE ST 2110 technology built around a centralized spine and leaf network architecture. Software-defined networking workflows were designed and implemented to orchestrate the routing of media flows over the network, which involves transporting heavy network Real-time Transport Protocol (RTP)[1] payloads. The system incorporates an extensive list of SMPTE ST 2110 endpoints, which include IP network encapsulation/de-encapsulation gateways, up /down/cross converters, IP-IP re-streamers, production switchers, audio mixing systems, replay capture and playout systems, graphics processors, clip players, and QC devices. The endpoints primarily leverage Networked Media Open Specifications (NMOS) IS-05[2] protocols for device discovery/registration and connection management.

PAC-12 Operations and Workflows

The PAC-12 Networks Engineering department is known within the broadcast industry for its innovative approaches to developing and implementing video and audio over IP signal transportation for HD broadcasts, a centralized production model for live remote events, also known as REMI, and cloud-based storage of media assets. Each season, PAC-12 Networks produces 850 live events, of which 700 are produced as

1545-0279/24©2024SMPTE

a Multicam or software-defined production. The PAC-12 Networks Multicam model leverages the PAC-12 private WAN to transmit camera feeds with audio, tallies, comms, and score/clock data from the venue to the production facility, allowing the director and technical director to make all the cuts. The Multicam units utilize J2K encoders/decoders, DSPs, tally controls, intercom systems, analog telephone adaptors (ATAs), etc., to facilitate this workflow. The Multicam model reduces the size of the mobile unit and crew required onsite while increasing production flexibility by allowing the traditional production crew to produce multiple events on the same day using the facility's standardized equipment.

The software-defined production model maximizes some of PAC-12 Networks best assets while remaining cost-efficient for smaller productions and events. Software-defined production utilizes commerical-off-the-shelf (COTS)-based hardware and an all-in-one production software solution to move the production capture hardware out of the mobile unit and onto the campus, where it connects to the PAC-12 private WAN. Software- defined production crews have all traditional broadcast features at their fingertips, accessing the production system remotely and combining several control room positions into 2-3 operators. The result is another technically innovative production that is agile and lightweight enough to adapt to changes in production plans, crew availability, technology, and budget.

> The software-defined production model maximizes some of PAC-12 Networks best assets while remaining cost-efficient for smaller productions and events. Software-defined production utilizes COTS-based hardware and an all-in-one production software solution to move the production capture hardware out of the mobile unit and onto the campus, where it connects to the PAC-12 private WAN.

Project Timeline and Key Technology Decisions

PAC-12 Networks—previously operating out of a facility in San Francisco since 2012—started to investigate transitioning to a new facility due to its lease terms set to expire in June 2023 at the previous facility. As such, PAC-12 decided in early 2022 to relocate to a facility located at the Bishop Ranch campus in San Ramon, CA.

As a result, it was also crucial to determine key technology solutions and providers to help PAC-12 build its state-of-the-art facility. With the ongoing supply chain issues and faced with the challenge of potential long manufacturer lead times, some key decisions and orders had to be made much earlier on in the process. After visiting the 2022 NAB Show and evaluating technology solutions with various vendors and colleagues in the industry, PAC-12 Networks decided to implement Imagine Communication's Magellan Control System and Selenio Network Processors built on top of Arista Networks' spine-and-leaf network architecture for their new build-out. The decision was based on successful use cases deployed with this general technology stack at other sporting venues and major sports networks.

Construction of the new facility in Bishop Ranch started in November 2022, and PAC-12 Networks had fully relocated into this facility by June 2023.

Broadcast Control System, Network Architecture Design and SDN

The broadcast technology infrastructure is built upon an underlying spine-and-leaf network topology utilizing Arista hardware supporting entire SMPTE ST 2022-7 redundant seamless protection switching (red/blue network) workflows. The network architecture includes:

- Two PTP Feeder Switches (Arista DCS-7020SR-24C2)
- Two Media Network Spines (Arista DCS-7804R3 with 400G and 100G interface line cards)
- Four Small AV Leaf Switches for 1G low-bandwidth audio devices (Arista DCS-7050TX3-48C8)
- Sixteen Medium Video Leaf Switches for 10/25G video devices (Arista DCS-7280SR3-48YC8)

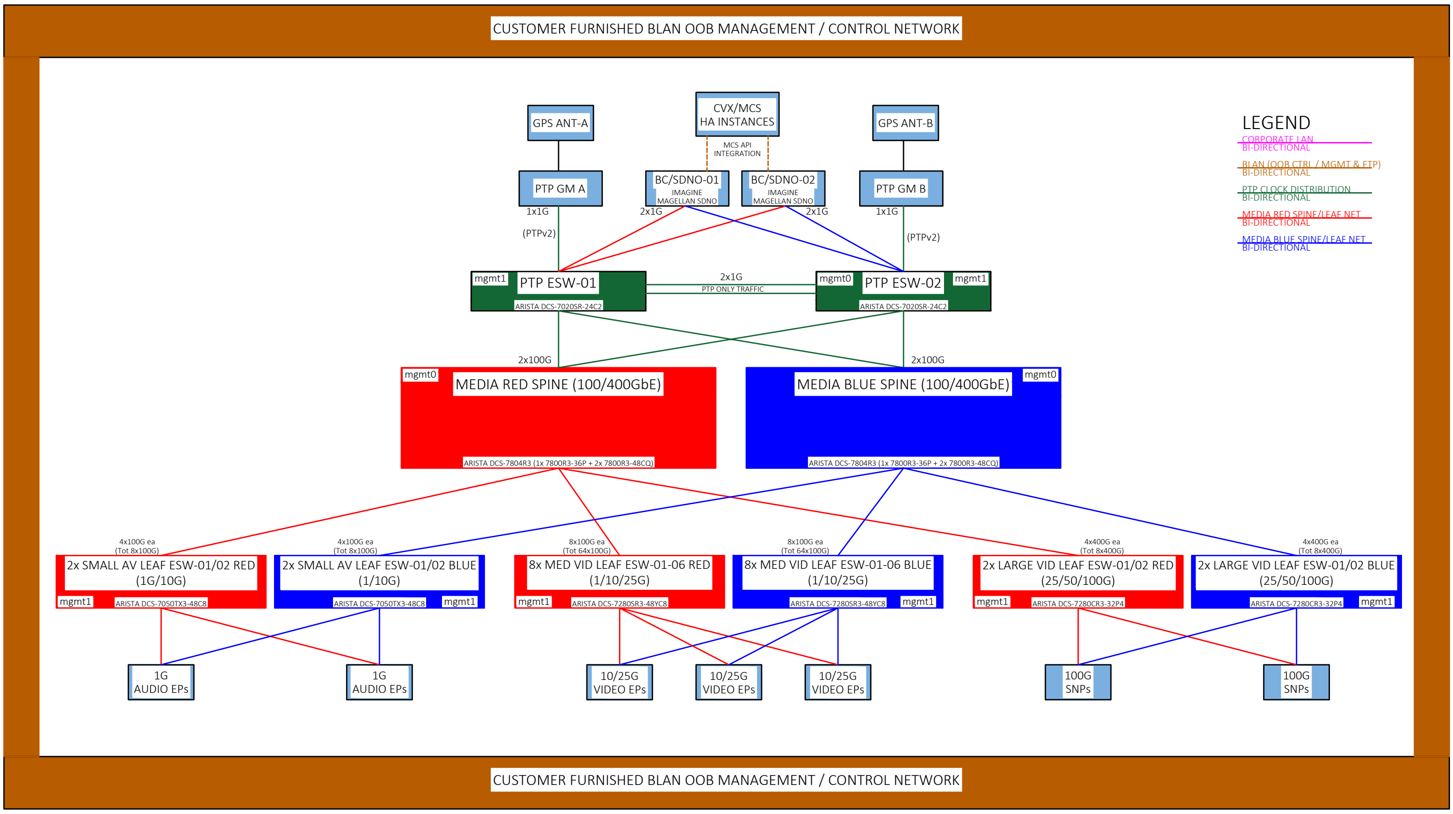


Figure 1. PAC-12 Networks SMPTE ST 2110 network topology.

- Four Large Video Leaf Switches for 100G processing devices (Arista DCS-7280CR3-32P4)

Figure 1 illustrates the overall SMPTE ST 2110 network topology.

It is important to note that the network architecture is fully routed, and most interfaces and network-network links are configured as Layer 3 routed ports with CIDR /30 subnets (Classless Inter-Domain Routing)[3] for point-to-point logical connections with the endpoints. Various topology segments are configured as switched Layer 2 VLAN interfaces with L2 uplinks to the spines terminating to Layer 3 SVIs (Switched VLAN Interface). We will discuss a particular use-case utilizing Layer 2 segmentation later in the paper.

All network uplinks from leaf switches to the spines are configured utilizing Border Gateway Protocol (BGP)[4] to facilitate unicast routing.

The routing orchestration and broadcast control layer functions are provided by Imagine's Magellan Control System. The control system applications are hosted on a redundant pair of hardware servers. The system also hosts the Registration Discovery Server (RDS) services and endpoint database, which consists of all the SMPTE ST 2110 devices that Magellan ultimately controls and their associated control protocols. These protocols combine natively used drivers by Selenio Network Processors and NMOS IS-05. All endpoint NMOS IS-04 discovery and registration API communication occurs with the Magellan RDS server. Magellan also communicates with NMOS capable endpoints over the IS-05 connection API and facilitates routing connection management messaging to receivers by patching the Session Description File (SDP) authored from the sender. All router database functions and configurations are managed under Magellan.

Magellan Control System interacts with the Arista Media Control Service (MCS) REST application programming interface (API). Magellan can manage and monitor real-time telemetry data for media within the underlying network topology through Arista MCS. As a result, MCS interprets user commands such as broadcast routing operations, which then works as an abstraction layer to create fast programming of static multicast routes into the network multicast forwarding information base (MFIB)5 tables of the physical switches. Additionally, MCS supports monitoring and controlling the reserved bandwidths of multicast streams over network links. This includes 3 Gbits/s for 1080p video flows, 12 Mbits/s for audio flows, and 5 Kbits/s for ANC data flows. This mechanism allows the network topology to be bandwidth-aware, enabling the networking architecture to be non-blocking. This is in contrast to using standard multicast routing protocols, such as PIM Sparse-Mode[6] and IGMP[7] in a SMPTE ST 2110 network. More on this topic will be discussed later.

The MCS integration with Magellan Control System allows the latter to control network traffic flows directly within the network architecture. This can only be accomplished after configuring all the Arista switches that have MCS enabled in Magellan with their corresponding management IP addresses and switch MAC addresses. Additionally, MCS sends notifications to Magellan to inform operators if a route has been successfully made or if a network path is impacted.

ST 2110 Networked Endpoints

Imagine's Selenio Network Processors (SNP) carries most of the media processing functionalities. Sitting at 1RU, these purpose-built hardware processors can be licensed to perform various broadcast-related functions. The primary functions of the SNPs at PAC-12 Networks include:

- SDI to IP encapsulation and IP to SDI de-encapsulation
- Frame-Synchronization
- Up/Down/Cross Conversion
- IP—IP re-streamers, allowing bridging of outbound essence RTP streams—with their unique multicast identification—and incoming IP essence streams from the same or disparate sources.
- Virtual re-entries, allowing the grouping of disparate video/audio/ANC essences and treating them as a virtual source & destination. This can also be thought of as a virtual patch panel.
- Multiviewer processing with metadata alarming.

Thirty-four SNPs were deployed for the build to support the functions listed.

Five Grass Valley K-Frame production control switchers controlled by Korona 3 ME (mixing effect) surfaces—one for each broadcast production control room—were deployed by PAC-12 Networks. Each switcher is capable of 48 IP inputs and 24 IP outputs.

In conjunction, five Calrec audio production/mixing systems were deployed with 48-fader Artemis control surfaces. Each surface supports up to 256 DSP input channels. Two redundant pairs of Impulse cores support the core routing. The first pair supports I/O for the two larger control rooms, and the second is used for the remaining three smaller control rooms. Each control room supports up to 256 audio flows, wherein each flow consists of 8 channels of 24-bit LPCM audio sampled at 48 KHz, with a packet-time of 1ms.

Five Ross Xpression graphics key/fill processing engines and five clip key/fill players were installed to support on-air graphics and clip playout requirements for each control room. An Evertz Dreamcatcher instant-replay capture and playout system was also deployed to support 26 ingest channels and 16 playback channels. Finally, SMPTE ST 2110-capable QC monitoring stations and tools were used for testing and troubleshooting. This setup included Telestream PRISMs, Wohler audio monitoring, and Plura video monitors.

PTP Considerations

For any SMPTE ST 2110 build, it is imperative to design and put in place a robust PTP distribution system across the media network. At PAC-12 Networks, a redundant set of PTP Grandmasters locked to GPS, was installed utilizing the SMPTE ST 2059-2 profile. All downstream network switches are configured in boundary clock mode with PTP priorities configured to provide maximum resiliency. The following PTP domain and messaging frequencies were used:

- Domain: 126
- Announce Interval: 0 (1 msg/sec)
- Sync Interval: -3 (8 messages/sec)
- Delay Request: -3 (8 messages/sec)

PTP synchronization is crucial to ensure the video/audio/ANC RTP essence streams are properly aligned. Each IP essence stream datagram must have an RTP time stamp in the RTP header field that is directly correlated to the PTP synchronization of the endpoint with the upstream master clock.

It is important to monitor TROffset values for ST 2110-20 video streams. TrOffset refers to the time difference between the PTP alignment point and the start of active video within a frame of video. For narrow-profile based senders (SMPTE ST 2110-21), this TROffset reading is approximately 623 microseconds for commonly used broadcast formats i.e. 720p/59.94, 1080i/59.94, 1080p/59.94 etc. This offset value directly correlates to the start of active video (SAV) within an SDI frame for the above formats. For example, SAV for 720p/59.94 is line 26, SAV for 1080i/59.94 is line 21, and SAV for 1080p/59.94 is line 42. All these line positions, depending on the format type, correlate to the same TROffset value of around 623us. It is important to note that most broadcast processing equipment capable of SMPTE ST 2110 usually supports narrow-profile based senders.

> FOR ANY SMPTE ST 2110 BUILD, IT IS IMPERATIVE TO DESIGN AND PUT IN PLACE A **ROBUST PTP DISTRIBUTION SYSTEM** ACROSS THE MEDIA NETWORK.

NMOS and Third-party Implementation

As SMPTE ST 2110 and NMOS capable endpoints (also referred to as nodes) are brought up online in the system, they will typically advertise their capabilities and resources under the Node API and post them to the Registration API hosted by the RDS server in the Magellan Control System. This mechanism is part of the NMOS IS-04 specification. IS-04 supports DNS-SD for nodes to discover the RDS server, automatically but in this application, the RDS server IP and TCP port (3214) were statically configured in the endpoints. Magellan utilizes an application, “Discovery Hub,” to automatically discover NMOS IS-04 capable endpoints as they are enabled.

The next part of the configuration process is associating each source and destination from a node with their corresponding IS-04 essence Global Unique Identifier (GUID) within the Magellan Control System’s database editor. This GUID is a string of unique characters identifying the node’s advertised senders and receivers. Typically, these GUIDs are persistent and will not change upon a physical endpoint reboot. As GUID associations are made, a friendly source or destination router mnemonic is assigned, i.e., CAM 11, PSWR 101, QCMON 1, etc.

A critical part of the database configuration involves specifying the physical network switch, interface type, and physical port to which the endpoint is directly attached. Depending on the essence type, the flow bandwidths are also specified, for example, 3 Gbits/s for video flows, 12 Mbits/s for 8-channel 1ms p-time audio flows, and 5k bits/s for ancillary data flows. As discussed, this is used for MCS to police network traffic across links.

> A CRITICAL PART OF THE DATABASE CONFIGURATION INVOLVES **SPECIFYING THE PHYSICAL NETWORK SWITCH, INTERFACE TYPE, AND PHYSICAL PORT** TO WHICH THE ENDPOINT IS DIRECTLY ATTACHED.

After registering the networked endpoints into the RDS server and programming their corresponding senders and receivers are programmed in the Magellan database, typical broadcast user actions such as routing a source to a destination can be easily performed. Magellan receives route commands from router control panels or other third-party crosspoint controllers, Magellan will perform NMOS IS-05 Session Description Protocol (SDP) patches into the receiving end-point’s transport_file field utilizing the Connection API. The transport_params field includes the sender’s source IP and destination multicast IP. It is important to note that the Imagine SNPs use their native protocols for all connection management.

Traditional SMPTE ST 2110 Workflows vs. Software Defined Networking ST 2110 Workflows

Traditional SMPTE ST 2110 workflows rely on standard multicast protocols such as Protocol Independent Multicast (PIM) and Internet Group Multicast Protocol (IGMP) to handle the overall forwarding of traffic across the networking infrastructure. In this environment, routing functions can be non-deterministic, and given that PIM by nature is not bandwidth-aware, there is a likelihood that network links across switches might become over-subscribed, leading to network congestion and packet drops, which can be undesirable in a live broadcast environment.

In an SDN environment, the following sequence of events occurs as route commands are completed (**Fig. 2**):

- Utilizing a router control panel or a third-party crosspoint controller, route requests are made using an underlying routing protocol such as LRC+ or SW-P-08.
- Magellan Control System interprets these route change commands for switching crosspoints from the source to the intended destination.
- Magellan will patch the SDP data from the sender to the intended receiver utilizing NMOS IS-05 connection API. In addition to the attributional data describing the essence within the SDP file, the sender’s source IP and destination multicast group address is also specified.
- In typical IGMP workflows, the receiver will interpret the source and destination IP from the SDP and issue an IGMP (S,G) or (*,G) Join for the multicast address depending on its support for source-specific multicast or any source multicast. The network topology will be responsible for forwarding the traffic. Essentially, the broadcast controller will treat the entire network topology as a black box without control over it.
- When utilizing SDN workflows—in addition to the SDP patch, Magellan will issue a POST over the appropriate REST API to MCS, which will, in turn, program static multicast routes into the Multicast Forwarding Information Base (MFIB) table of the network switch that is directly connected to the destination endpoint.

When static multicast routes are created, the SMPTE ST 2110 RTP flows are pushed to the intended receiver. MCS will inform Magellan if the route was successful or not.

Arista CloudVision Portal (CVP) is utilized in the background, which gathers real-time network streaming telemetry data from MCS and the network switches to provide network monitoring and diagnostic functions.

Critical Lessons Learned

Third-Party Control System Issues

PAC-12 Networks utilizes a router control system and panels that directly interface with the Magellan Control System over the LRC+ communication protocol. This system runs on a cluster of three compute virtual machines and uses multiple services that handle different tasks. One such service is the Metadata service, which is responsible

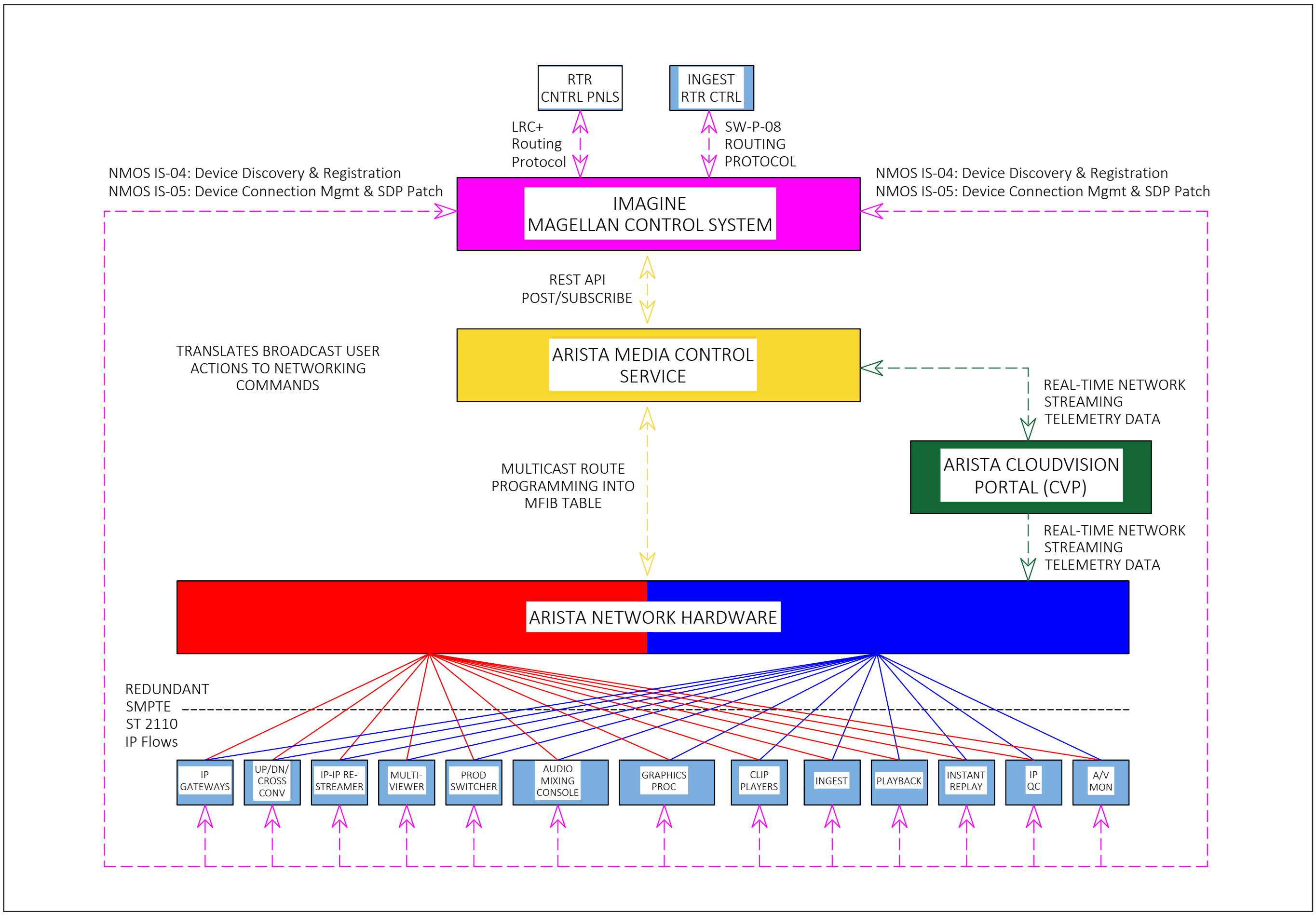

RTR CNTRL PNLS
INGEST RTR CTRL
LRC+ Routing Protocol
SW-P-08 ROUTING PROTOCOL
NMOS IS-04: Device Discovery & Registration
NMOS IS-05: Device Connection Mgmt & SDP Patch
IMAGINE MAGELLAN CONTROL SYSTEM
NMOS IS-04: Device Discovery & Registration
NMOS IS-05: Device Connection Mgmt & SDP Patch
REST API POST/SUBSCRIBE
TRANSLATES BROADCAST USER ACTIONS TO NETWORKING COMMANDS
ARISTA MEDIA CONTROL SERVICE
REAL-TIME NETWORK STREAMING TELEMETRY DATA
MULTICAST ROUTE PROGRAMMING INTO MFIB TABLE
ARISTA CLOUDVISION PORTAL (CVP)
REAL-TIME NETWORK STREAMING TELEMETRY DATA
ARISTA NETWORK HARDWARE
REDUNDANT SMPTE ST 2110 IP Flows
IP GATEWAYS
UP/DN/ CROSS CONV
IP-IP RE-STREAMER
MULTI-VIEWER
PROD SWITCHER
AUDIO MIXING CONSOLE
GRAPHICS PROC
CLIP PLAYERS
INGEST
PLAYBACK
INSTANT REPLAY
IP QC
A/V MON

Figure 2. PAC-12 Networks BC system architecture.

for informing other services about router crosspoint status changes. During the early stages of the project there was an issue that occurred. Whenever a route change was triggered through this system (via hard-panel or user interface [UI]), multiple crosspoint requests, instead of one, were sent to a specific set of third-party NMOS IS-05-capable IP de-encapsulators from a different manufacturer. Due to the slightly slower nature of this vendor's NMOS IS-05 connection mechanism, the northbound third-party router control panel system interpreted it as a failed route. As a result, the system triggered approximately 7-8 crosspoint route change commands serially, resulting in a slow switching response (~15 sec) in some third-party NMOS IS-05 capable receivers. This issue was addressed in a firmware update; however, it resulted in a much more critical problem with the way the hardware panels and UI panels communicated. Each button push essentially informed every panel of this state change in the network, an undesirable behavior referred to as a broadcast storm. As more users activated route changes through a button press, multiple messages attempted to sync with the Metadata service. During peak situations, the Metadata service handled hundreds of thousands of requests, causing it to crash, rendering the router panel and UI system non-functional. After working tirelessly with the vendor to monitor the issue and identify the root cause, the vendor was able to resolve this issue with a new software patch. Now, the hardware panel service and the UI panels send out the crosspoint route request message only once, and switch response times are significantly faster.

Audio System Frame Redundancy

PAC-12 Networks implemented an audio routing core system manufactured by Calrec Audio. The system architecture comprises redundant Impulse routing core frames. This active standby system architecture provides a quick failover option to the redundant frame with little to no interruption to on-air activities if the primary frame fails. This type of redundancy and fault-tolerant architecture is crucial to any live television network operations.

Furthermore, SMPTE ST 2110-30 audio stream-level redundancy is supported by redundant ST 2022-7 interfaces.

One method to implement this frame-level redundancy requires that the corresponding ST 2110 ports across both primary and redundant core frames share a virtual unicast IP so that the sender's source IP will technically not change due to a frame failover and would always be represented by the virtual IP. This meant that for each ST 2110-30 interface of the primary/redundant system pair, there would be a total of three unicast IPs: one physical IP for the primary frame interface, one physical IP for the redundant frame interface, and one virtual IP shared between both primary and redundant interfaces. The corresponding destination multicast addresses of the senders would be shared across both frames.

This approach would not have worked with the original networking topology configuration wherein every interface was configured as a Layer 3 routed interface. Magellan requires a single physical or virtual switch interface in its configuration of senders and receivers to talk with Arista MCS. Since the primary and redundant audio frames were physically connected to separate network leaf switches, the Layer 3 CIDR /30 schema would not be functional.

As a result, it was decided to re-configure the audio system to be provisioned on its own Layer 2 VLAN on a CIDR /24 subnet with Layer 2 uplinks configured as port channels from the red/blue leaf switches to the spines. A Layer 3 routed SVI[8] was configured at the spine to allow routability of audio streams to and from the rest of the system. Subsequently, the sources and destinations from the audio system would be configured in Magellan to point to this Layer 3 SVI instead of a physical interface.

Additionally, it was noted that the audio systems' NMOS IS-04 node API, under /*self*, exposed the physical IP of the active frame instead of the virtual IP. The SDPs of the sender streams, however, did accurately reflect the virtual IP as the source IP. The resultant effect was that any receiver would not properly receive the audio streams due to the mismatch in source IPs. To address this issue, the audio sender sources in the audio system were configured in Magellan as unmanaged devices, instead of NMOS devices. This was done by manually entering stream metadata that included the source IP of the sender along with the virtual IP. This configuration allowed Magellan to "synthesize" the SDP with matching source IPs and virtual IP. The only caveat to this approach is that if any inherent stream configuration changes occur at the audio system (for example, changes in p-time or audio channel count), they would not be tracked by the manually authored SDP. However, this is not likely to occur since the audio formatting and parameters are locked and not expected to change. A firmware update is being developed that will address this issue.

Troubleshooting with Arista MCS API

Although there were no critical issues with the Arista MCS system, it's worth noting that the troubleshooting and diagnosis process for media-related networking issues in this type of software-defined networking environment is different. Instead of relying solely on the networking command line interface, PAC-12 Networks' engineering staff had to become familiar with the various MCS API commands to assist them with the testing and troubleshooting.

Conclusion

This paper described the build-out of PAC-12 Networks' new technology center utilizing SMPTE ST 2110 and SDN technology. The key system components for the facility and critical technology issues faced along with implemented solutions were detailed. The implementation phase for the project occurred over an extremely aggressive timeline to meet the collegiate sporting event start date. The PAC-12 Networks became fully operational for the season after the successful completion of the project in September 2023, enabling the production of 850 live sporting events.

Acknowledgments

In connection with the design and construction of the new PAC-12 Networks facility, special thanks are extended to PAC-12 Networks' team members: Hieu Ho, Howie Chung, Roman Kayutkin, Megan Kappus, Brett Derham, Elida Errante, Ryan Currier, Richard Katz, Larry Meyers. Former Pac-12 Engineering team members: Scott Heitman and Nick Brandt

Special acknowledgments are also extended to ASG's core engineering and project team members: Nik Kumar, Spencer Roberts, Clem Phillips, Nand Ganesh, Joel Johnson, Steve Young, Harrison Gardanier, Shaun Halladay, Mike Surman, Sam Craig, Matt Minihan, Michele Ferreira, Garrick Huey, and Dave Van Hoy.

The project would not have been possible without the support from our vendors and key partners. From Imagine Communications, special thanks go out to Jimbo Haneklau, Jay Chambers, Todd Barkes, Konrad Duda, and Steven Yang. From Arista Networks, thanks are extended to Ryan Morris, Sakti Arunachalam, Lesley Lurie, and Hytham Abu Safieh. Additional thanks are extended to our vendor partners supporting other critical areas of the workflows: Pat Schnieder (Calrec Audio), Dave Barnhardt (Evertz), Gaylen Clutter (Evertz), Joe Sgherza (Grass Valley), Brett Falcetto (Ross), and Sal Paglia (RTS).

References

1. Real-time Transport Protocol (RTP. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc3550
2. Networked Media Open Specifications (NMOS). [Online]. Available: https://specs.amwa.tv/nmos/
3. Classless Inter-Domain Routing (CIDR). [Online}. Available: https://datatracker.ietf.org/doc/html/rfc4632
4. Border Gateway Protocol (BGP). [Online]. Available: https://datatracker.ietf.org/doc/html/rfc4271
5. Multicast Forwarding Information Base (MFIB). [Online]. Available: https://www.arista.com/en/um-eos/eos-multicast-architecture
6. Protocol Independent Multicast (PIM). [Online]. Available: https://www.ietf.org/rfc/rfc2362.txt
7. Internet Group Management Protocol (IGMP). [Online]. Available: https://datatracker.ietf.org/doc/html/rfc3376
8. Switched VLAN Interface (SVI). [Online]. Available: https://www.ietf.org/rfc/rfc2674.txt

About the Authors

Nik Kumar is a broadcast solutions architect with over 17 years of experience leading engineering and integration teams to effectively complete system integration projects for customers in the broadcast, media, and entertainment industries. Kumar is currently Director of Engineering for Systems Integration at Advanced Systems Group (ASG)

Hieu Ho is a broadcast engineer who has over 12 years of experience in leading and managing maintenance engineering teams at various college and professional sports & news network facilities. Ho is presently serving as a lead broadcast technology engineer at PAC-12 Networks in San Ramon, CA.

A contribution received from the authors, January 2024. Copyright © 2024 by SMPTE.

DOI: 10.5594/JMI.2024/IIOY7619
Date of publication: 1 April 2024

1545-0279/24©2024SMPTE

Standards Technology Committee Meetings

On a quarterly basis, the Standards Community convenes for week-long TC Meetings. During these sessions, participants provide updates on progress and collaborate on advancing standards work.

3-5 June 2024
OTTAWA, CA

18-20 Sept. 2024
GENEVA, CH

Interested in hosting a TC Meeting?

KEYWORDS LONG-TERM PRESERVATION // METADATA // DCP // FILE FORMATS // CPP // DIGITAL PACKAGES // METS // EBUCORE

EN17650—The New Standard for Digital Preservation of Cinematographic Content

By Siegfried Fößel, Heiko Sparenberg, Nikolai Belevantsev, and Yi Lou

Abstract
With the cessation of analog film as distribution medium for cinematographic content, its use as a storage medium was also abandoned. This created new challenges in the long-term preservation of cinema movies—now in its digital form. The Academy of Motion Picture Arts and Sciences (AMPAS) drew attention to this in the two reports "The Digital Dilemma" in 2007 and 2012 and the International Federation of Film Archives (FIAF) gave feedback in "The Digital Statement." While storage capacities have continued to grow and are no longer the biggest problem, the increasing variety and also the rapid obsolescence of digital formats remains a hot topic. For this reason, the new European standard EN17650 has been developed to address this issue. This paper presents the requirements, considerations, and solutions discussed during the standardization project. It explains the system architecture, formats, and elements of a "preservation package" and the criteria for selecting specific metadata within the new standard.

In 2007, the Academy of Motion Picture Arts and Sciences (AMPAS) published "The Digital Dilemma" report,[1] which detailed the challenges facing the movie industry in the transition from analog film to digital data. An update of the report was produced in 2012.[2] Meanwhile, some problems have decreased due to technological advances and workflow changes. For example, storing substantial amounts of data from a digital representation of a movie or transmitting a Digital Cinema Package to a cinema is no longer a problem. The International Federation of Film Archives, French: Fédération Internationale des Archives du Film (FIAF) has also addressed this problem and provided information on the website in form of "The Digital Statement."[3]

Some of the challenges, however, remained or even increased, especially in the long-term preservation process. The number of digital formats and codecs is constantly increasing, other formats are becoming obsolete, and repositories in archives are used for publishing functions or artificial intelligence (AI) analyses. As a result, managing digital film archives and using metadata is becoming increasingly important. To ensure the preservation and accessibility of digital cinema content, the European Union (EU) included measures in its Rolling Plan for Information and Communication Technologies (ICT) standardization as early as 2017.[4] At the end of 2022, these actions were implemented in the European Committee for Standardization, French: Comité Européen de Normalisation (CEN) Technical Committee 457, resulting in two new standards, EN 17650:2022[5] and CEN/TR 17862:2022.[6] These two standards define a framework for preserving digital cinema content and its metadata, and many European archives were involved in this standardization process.

The intention of the standard was not to define another digital format but to reuse and extend existing formats to form a basis for organizing the content. For this reason,

1545-0279/24©2024SMPTE

metadata schemes such as METS[7,8] EBUCORE,[9] PREMIS[10] or SMPTE IMF playlists[11] were added and extended to allow a rich data set. Guidelines are provided as to which standard formats should be stored additionally in case proprietary formats need to be archived.

"The Digital Workflow" chapter of this paper describes the changes in the movie production workflow, "The Placement of Standard in the Archiving Context" explains where the new standards fit in the world of file and essence formats. "The Cinema Preservation Package" chapter provides a detailed description of the structure and elements of the preservation package. In the "HFF Pilot Project," the integration of the Cinema Preservation Package in the working environment of the University of Television and Film Munich (HFF) film school is presented as an example. The remaining chapters conclude with results and a summary.

> The Cinema Preservation Package (CPP) contains all data of a cinematographic work in one package and can be used for long-term preservation of digital film assets. It is compliant with the Open Archival Information System (OAIS) principle. Subpackages can be separated to share only part of the work. Multiple subpackages can be logically grouped into a collection without the need for data duplication.

The Digital Workflow

The advent of digital cinema changed not only the distribution and the projection/display of cinematographic works but also the entire production workflow and, with it the types of preservation objects. Cinematographic works are now fully digital and are no longer recorded, stored, and vaulted on analog film. Film assets are also reused today for AI processing, remastering, and, in the future, for digital asset libraries, where film producers can adopt elements of the film for new projects.

These changes arrived with a manifold of new digital formats during production and mastering. The Digital Cinema Package (DCP)[12] is a stable format for the distribution to the cinema and the equivalent to the film distribution reel. The Interoperable Master Format (IMF)[13] is an accepted master exchange format, but a common standard or framework for the preservation of assets used during production, mastering or in general around a digital cinematographic work did not exist. This was the motivation in 2014 to propose a new standard for preserving cinematographic works. The proposal was added to the "Rolling Plan for ICT standardization" and granted as a new CEN project in 2019. Work on the standard started at the end of 2019 with a small project team under the supervision of CEN TC457[14] and is now published as European Standard EN17650.

Digital workflows in the media industry are rapidly embracing new technological advances and frequently changing the file formats used. This stands in contrast to the needs of archivists for a long-term preservation format, description, and structure. The question was how can digital formats for cinematographic works be stored so that they can be reliably interpreted even in 10, 20, or 50 years from now? The new standard tries to answer this question and offers methods to help in this preservation task. It does not answer the question of which storage medium should carry the files. It also does not prevent the content owner from storing the assets in formats that are not suitable for long-term preservation.

Placement of the standard in the archiving context

To better understand the application field of the new standard, the International Organization for Standardization/Open Systems Interconnection (ISO/OSI) layer model for network communication is transferred to the field of long-term archiving of digital movies. It uses an abstraction of the seven-layer ISO/OSI model and applies it to the long-term archiving area (**Fig. 1**).

The first three layers are media-dependent and define how data is stored and organized on a specific medium. An example is:

- Layer 1, where the bits are physically written on a data carrier.
- Layer 2, where the data is grouped in blocks, e.g. in 512- or 4096-byte blocks for a sector on an HDD or in 480KiB blocks on linear tape open (LTO).

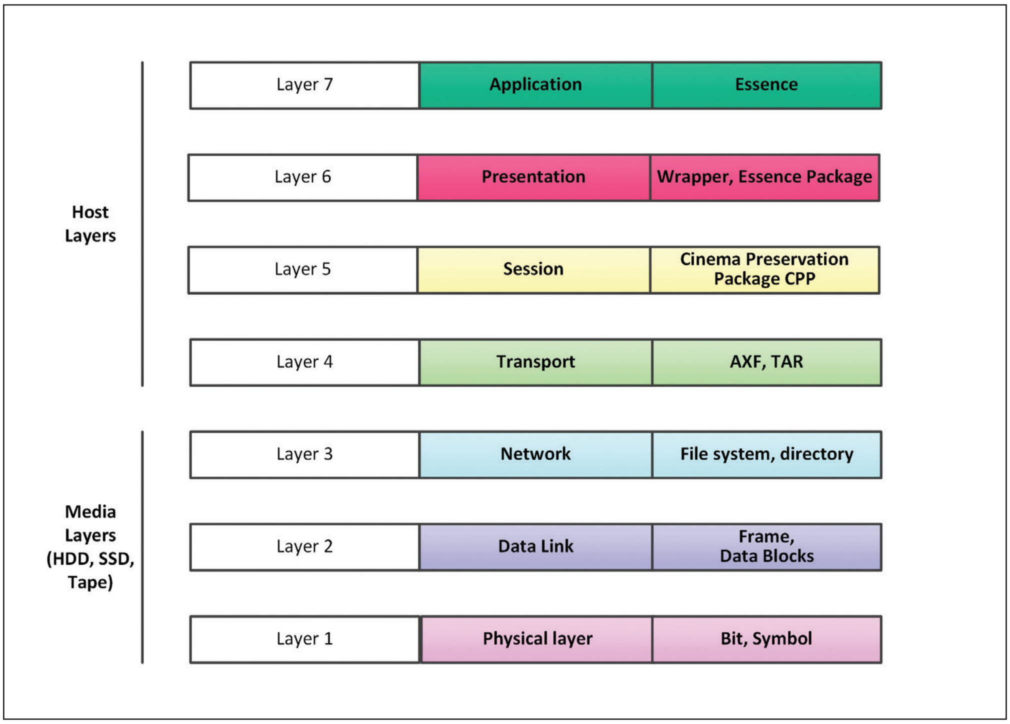


Figure 1. Layer model for long-term archiving.

- Layer 3, where the blocks are grouped and organized together, e.g., in a file system like ext4 or LTFS.

The three lower layers define how data is organized and are not dependent on the type of data. The higher layers are more dependent on its use case. Layer 4 bundles the data and allows an easy migration or transfer between media carriers. Examples of such formats are AXF (Archive eXchange Format) or Tar (archive file format, name is derived from tape archive). In this paper, we discuss layer 5, which collects all data of a cinematographic work in one package, the so-called Cinema Preservation Package (CPP). Part of the CPP can be any audiovisual file in addition to metadata and helper files. Layer 6 contains the typical audiovisual files like MP4 and a multi-file representation format, such as DCP or IMF. Layer 7 is then the pure essence.

The Cinema Preservation Package (CPP)

Development Goals of the New Standard

The main goal of the new European Standard was to define a way to package all elements of a cinematographic work for digital preservation. It should reuse existing standards as much as possible. The structure and content of the package should be human-readable or understandable without relying on extensive computer analysis programs. Therefore, converting the filenames to abstract Universally Unique IDentifier (UUID) names, flattening the folder structure, and packaging all files into a tar file was not an option.

The standard should define methods to store content in physical and logical structures and describe relationships and metadata for its components. The definition of logical structures allows the description of virtual packages (collections), like a delivery to a film festival or a set of files from a specific scan, without storing the data multiple times or in a separate folder.

The CPP should use and extend existing XML schemes to store structural, descriptive, technical, and provenance metadata. The aim was not to define another metadata standard but to use existing standards to advantage. In practical scenarios, storing metadata in an online database is common. However, for long-term preservation, it is recommended to include a copy of the metadata alongside the content, following the example set by the Open Archival Information System (OAIS).[15] This approach ensures that the package remains self-contained. For this purpose, a synchronization mechanism between online- databases and XML files might be necessary.

Structure

The first requirement for the preservation standard was the definition of a clear package structure that allows storing the content in a well-defined and human-interpretable physical order while also providing the option of forming logical collections. To allow combining and exchanging only parts of the cinematographic work, sub packages for parts of the work were defined in the full package (**Fig. 2**). Each sub package contains only one type of essence in the data folder, technical and provenance metadata for this essence are defined in the metadata folder and helper files for this essence in the ancillaryData folder. The type of content or data files can be an image sequence (in an imagePackage), sound files (in a soundPackage), timed text files (in a timedTextPackage), composed files like MP4 files (in an audiovisualPackage), a

1545-0279/24©2024SMPTE

set of files belonging together as a package like DCP, Digital Production Partnership (DPP) or IMF package (in a componentizedPackage) or extra files (in an extraPackage).

The content of such subpackage is described in an XML file. This was achieved by using and adopting the METS schema. The Metadata Encoding & Transmission Standard (METS) file (named packingList.xml in the <sub-package> folder) links provenance and technical metadata to content files in the data folder. Multiple subpackages together with further metadata, ancillaryData and playLists form the full package. Here in the root folder the same principle is used as in the subpackages. A METS file (named preservationPackingList.xml) links descriptive and provenance metadata to the METS files in the subpackages. The METS file in the subpackage corresponds to the data files in the subpackages. A hierarchical system of METS files (kind of inventory list) exists.

Virtual packages can be defined in the preservationPackingList.xml file. An example is the elements from a scan (**Fig. 3**, SCAN_20200315). The structMap element in the XML files references the METS files in the related subpackages (**Fig. 4**).

The reason for the organization in sub packages was also because in daily operations only parts of a cinematographic work are processed or moved. In consequence, not all parts of the CPP should be touched or scanned when a content element is changed, added, or removed. The re-generation of hash values can especially lead to time-consuming workloads.

The METS files are used to list all files with their checksums, create logical and physical views on the content, and link data files to other metadata files. Technical, descriptive, and provenance metadata are not embedded directly in the METS files. One reason is that METS itself does not always offer good metadata descriptions, and the option to embed other schemes into METS files

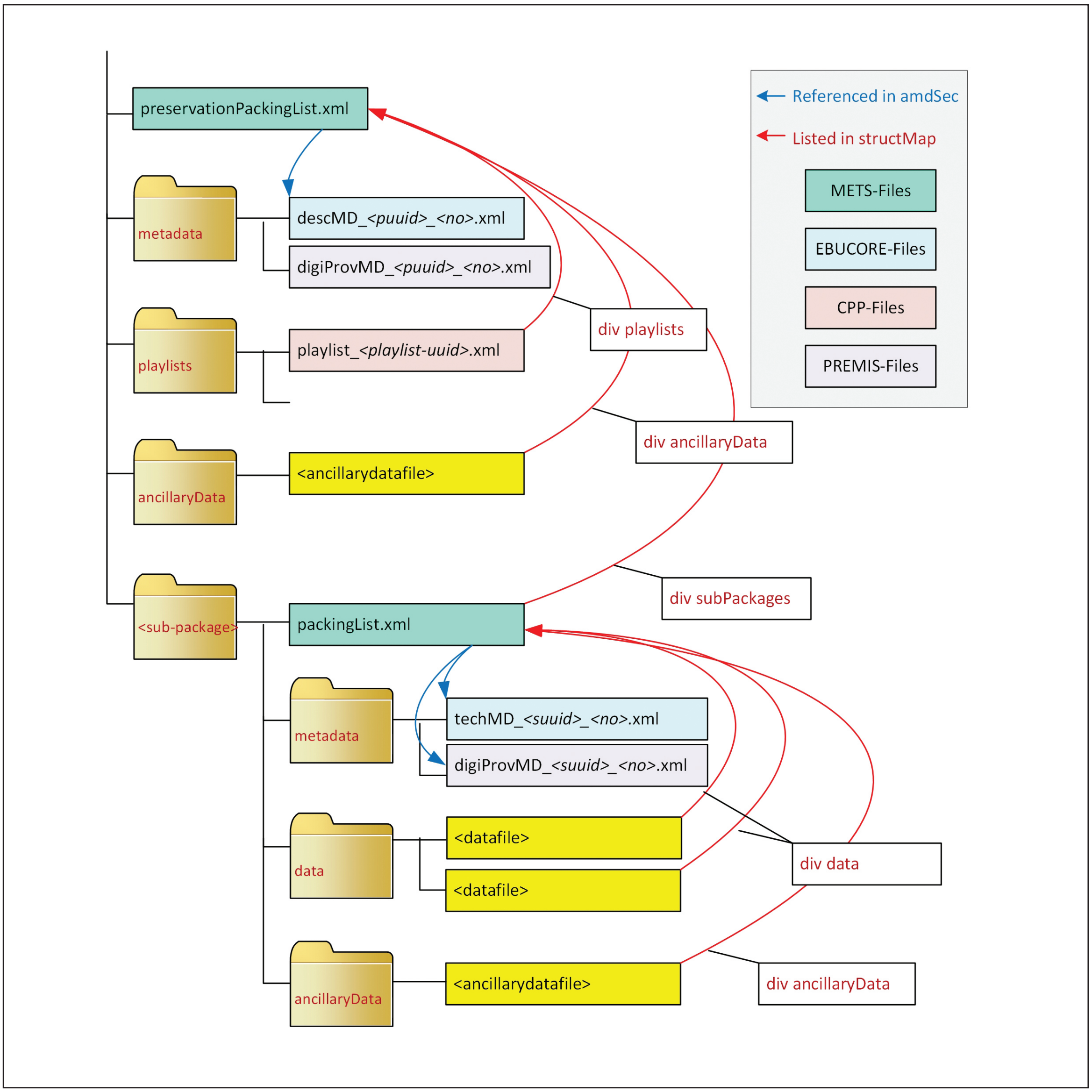


Figure 2. Basic structural elements of a CPP.

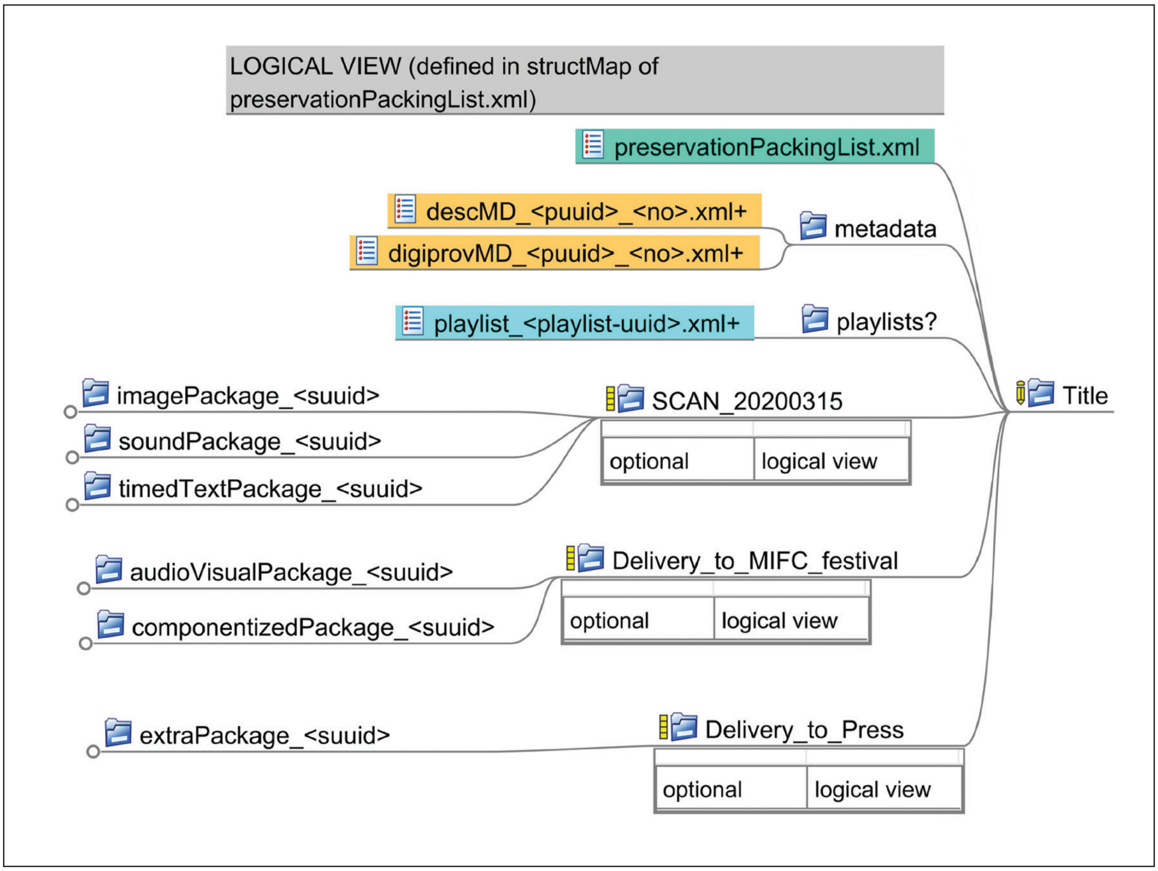

LOGICAL VIEW (defined in structMap of preservationPackingList.xml)
preservationPackingList.xml
descMD_<puuid>_<no>.xml+
digiprovMD_<puuid>_<no>.xml+
metadata
playlist_<playlist-uuid>.xml+
playlists?
imagePackage_<suuid>
soundPackage_<suuid>
timedTextPackage_<suuid>
SCAN_20200315
optional
logical view
Title
audioVisualPackage_<suuid>
componentizedPackage_<suuid>
Delivery_to_MIFC_festival
optional
logical view
extraPackage_<suuid>
Delivery_to_Press
optional
logical view

Figure 3. Logical views (collections) within a CPP.

increases the complexity of schema checking and complicates manual editing.

An additional feature at the root of a CPP is the utilization of playlists within the playlist folder, which enables the generation of pre-composed content. This is not only used for a playable piece of content e.g., a composition of image sequence and sound files but can also be used simply for a composition of files belonging together. The playlist binds individual data files together whereas the logical view in the METS preservationPackingList.xml file binds subpackages together.

Metadata Schemes

As a result of the investigation of metadata standards, the following formats are used in the CPP:

- METS for structural metadata (also virtual packages), hash values, and linkage of files
- EBUCore for technical metadata
- EBUCore for descriptive metadata (based on EN15744[16])
- PREMIS for provenance metadata

All proposed metadata standards could be adopted with their existing schema descriptions and inherent extension mechanisms. Therefore, no new schema definition was necessary.

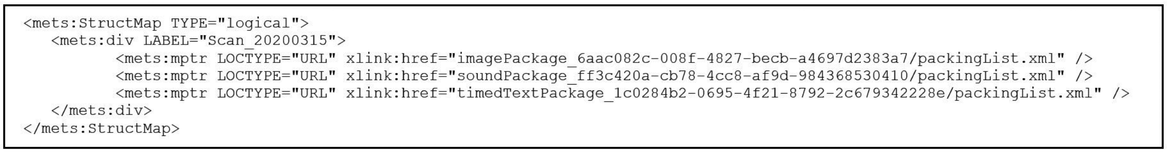
EBUCore was selected for the technical metadata description for two reasons. First, it already offers a large set of technical metadata parameters for the video and movie industry. In addition, some tools exist, like MediaInfo, that can automatically generate basic technical metadata files. Second, EBUCore has an easy extension mechanism that allows adding new metadata parameters without redefining the standard. In the standard a full set of technical metadata parameters are defined together with their mandatory or optional availability.

For the descriptive metadata, the selection was not so obvious. With EN15744, a minimum set of metadata for cinematographic works already exists. However, the standard describes only the data elements as ontology but not the language or syntax for implementation. This led to different implementations in the film archive world. The lack of examples also led to confusion and differences in applications. During the CEN project, the Deutsche Kinemathek worked with the Zuse Institute in Berlin on a new EN15744 implementation based on EBUCore for their internal purposes. This implementation was contributed to the project team that adapted it in a more generalized way for the CPP.

Identifiers (IDs)

METS uses many internal IDs for referencing and cross-linking of elements. For a CPP, however, also, global IDs are beneficial, especially in making such IDs available to archive management systems. In **Fig. 1**, the use of UUIDs is shown in the filenames as puuid (preservation package UUID) or suuid (sub-package UUID). In addition, the subpackage name contains the suuid. This allows easy identification of identical subpackages or preservation packages in the archive sys-

1545-0279/24©2024SMPTE



```
<mets:StructMap TYPE="logical">
   <mets:div LABEL="Scan_20200315">
          <mets:mptr LOCTYPE="URL" xlink:href="imagePackage_6aac082c-008f-4827-becb-a4697d2383a7/packingList.xml" />
          <mets:mptr LOCTYPE="URL" xlink:href="soundPackage_ff3c420a-cb78-4cc8-af9d-984368530410/packingList.xml" />
          <mets:mptr LOCTYPE="URL" xlink:href="timedTextPackage_1c0284b2-0695-4f21-8792-2c679342228e/packingList.xml" />
   </mets:div>
</mets:StructMap>
```

Figure 4. Description of virtual package in METS file.

tem and their related metadata. Each time the data elements change, a new UUID is generated.

Data formats

The CPP so far is agnostic to the used content formats in the data folder. However, recommendations for formats are listed in the standard. In general, well-defined formats should be used, such as ISO, International Telecommunication Union (ITU), SMPTE, Internet Engineering Taskforce (IETF) standards, etc., that can be referenced in the metadata files and for which open software implementations exist. Special placeholders in the metadata files are available for these references. If proprietary formats must be stored and archived, a description of the format should be stored together with the data files in the related ancillaryData folder.

The HFF pilot project

Figure 5 shows how the CPP is integrated into a large context/environment at the HFF film school in Munich. The CPP is the archiving format for the assets and is linked to different management systems.

The primary goal of the pilot project is to create a digital archive for the long-term preservation and efficient management of film assets produced by HFF students. Implementing the digital archive aims to achieve several key objectives:

The HFF long-term archive will function as a central hub where all film data generated since 1967 can be securely stored in one location. Existing descriptive metadata will be migrated from diverse databases to the long-term archive, ensuring that the information about the HFF's film heritage is catalogued to a uniform standard. Film assets within the archive are linked to their corresponding metadata, providing an overall view of entire collections and facilitating efficient search, access, and distribution processes.

An automated workflow for archiving new films must be established at the HFF. Students use the production manage-

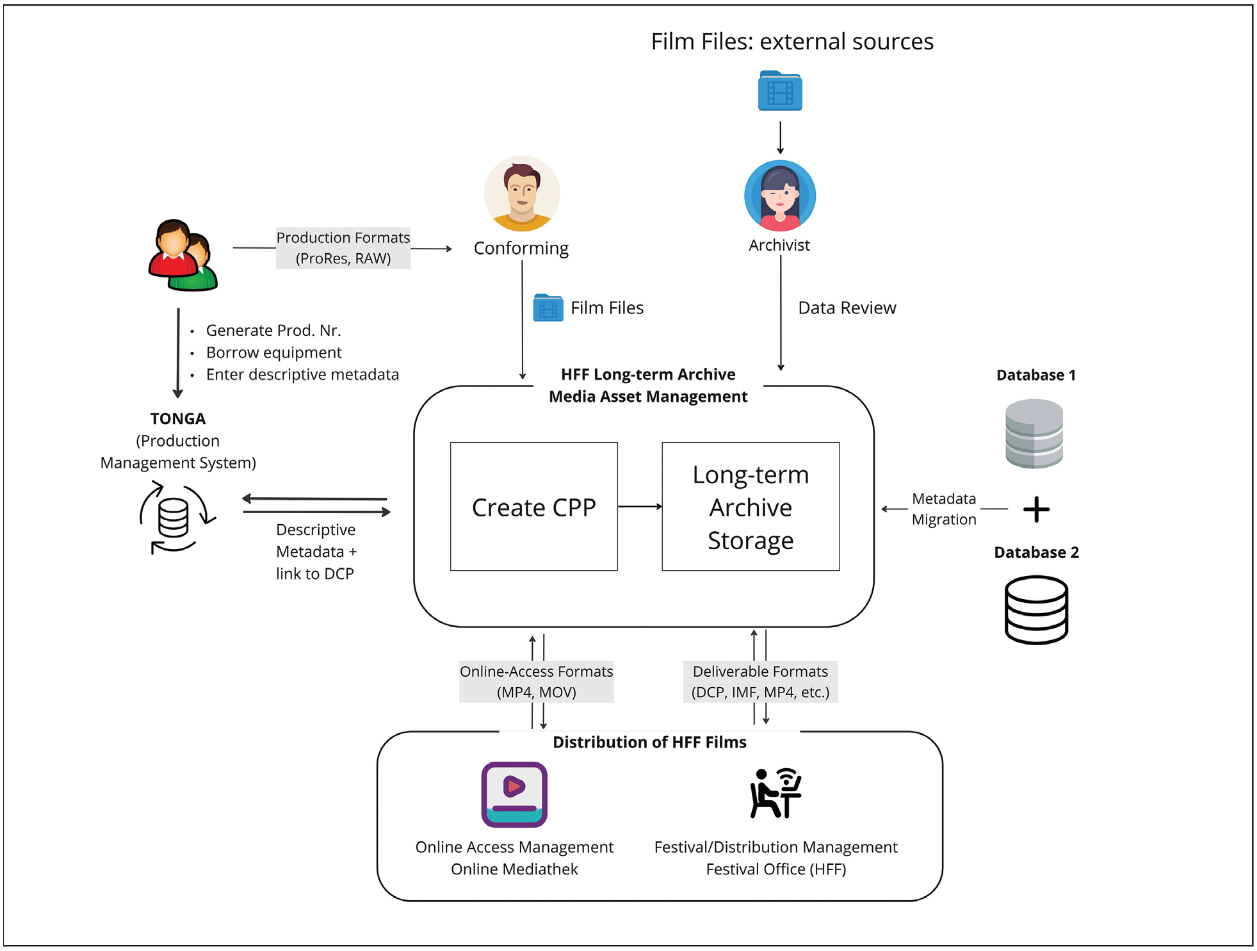


Figure 5. Example environment at HFF.

1545-0279/24©2024SMPTE

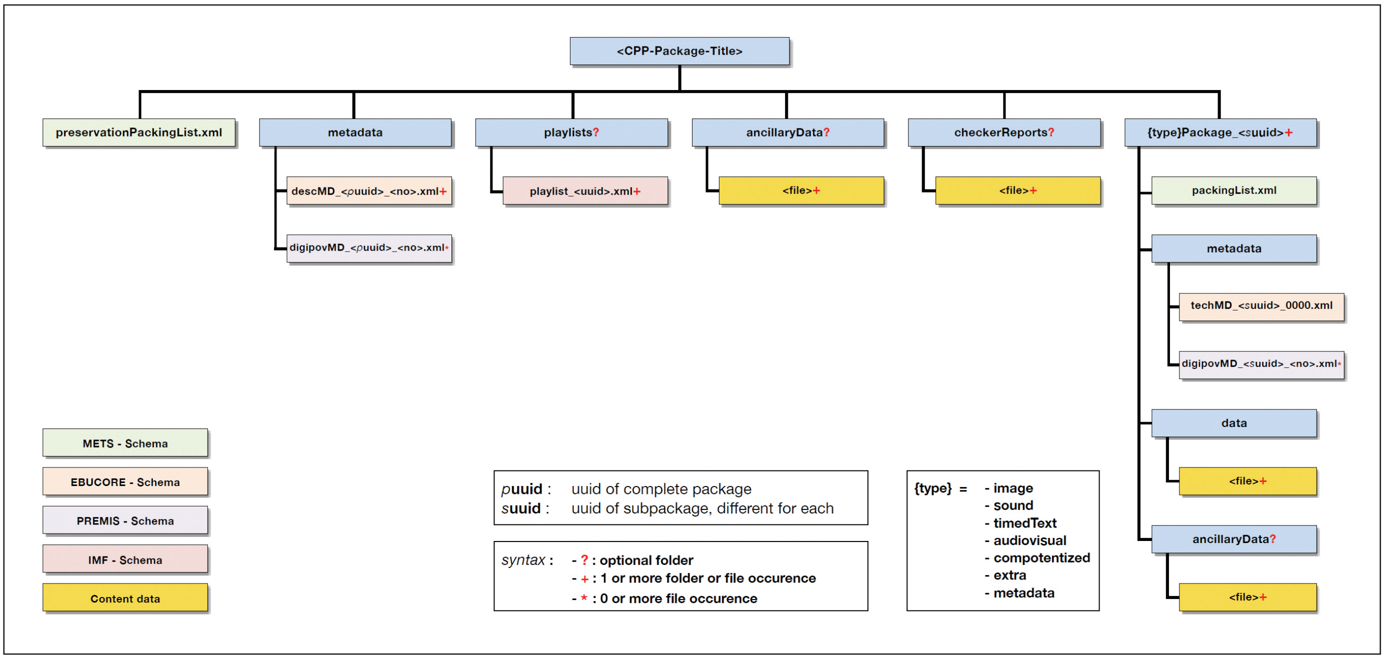


Figure 6. All components of a Cinema Preservation Package.

ment tool TONGA to enter descriptive metadata for their films during film production. Once film production is completed, the metadata will be automatically transferred from the production management tool to the long-term archive. The long-term archive incorporates features for ingesting film data, and all data for a film work should then be compiled into one CPP. The CPP is subsequently transferred to the storage location, ensuring the integrity of the archived content.

The user interface of the archive allows authorized users to explore and retrieve film assets stored in the system. These film assets can be retrieved based on different needs and reused for AI processing or other research projects that contribute to the innovation of new film technologies and a deeper understanding of film aesthetics. Additionally, linking the long-term archive to the distribution management system and the forthcoming online mediatheque can maximize the value and impact of HFF films. The long-term archive also stores information about usage rights, licensing agreements, and permissions associated with film assets, ensuring compliance with legal and copyright regulations, and preventing unauthorized usage when distributing films to various platforms.

Results

At the beginning of the project, the focus was more on the physical structure of the preservation package. During the project, archives requested to add logical views and provide more guidance on the use of the metadata. Especially for the technical metadata, many new metadata parameters were added to the standard. A complete overview of components in a Cinema Preservation Package is shown in **Fig. 6**. It also adds an optional folder called checkerReports. This folder is reserved for post-analysis files after the CPP is created, e.g., checks if hash values are still correct, schema checks, or consistency checks. These files are not included in the file lists inside the METS files, and the folder is considered as a reserved storage space for such files.

The result of the work is a complete standard for describing a Cinema Preservation Package.[5] A related Technical Report[6] gives additional information and guidance on implementing the standard and offers additional explanations to the structure. Various content combinations are described as references for concrete implementations in the technical report.

Figure 6 also shows all components of a CPP together with the used schemes for the metadata. Actual work is ongoing with further developing an open-source reference software that will enable a better understanding of the standard.

Conclusion

The reuse of existing standards and metadata schemes led to many discussions inside the technical committee TC457, which is responsible for this project, and the project team. Further discussions continue in an online forum of the Academy of Motion Picture Arts and Sciences.[17] In the end, it needed a lot of insight into existing standards and schemes, as the specific requirements created complex dependencies. However, with the newly proposed standard, a satisfactory solution was found that allows for easy future extensions and builds upon existing formats in the archive domain.

Acknowledgments

The authors thank the European Commission for the project SA/CEN/2016-05, "Preservation of digital cinema," in which the documents EN17650 and TR17652 were developed. The recent part of this work at HFF is funded by the German Federal Ministry of Education and Research under the funding code 13IHS220. Responsibility for the content of this publication lies with the authors.

1545-0279/24©2024SMPTE

References

1. Academy of Motion Picture Arts and Sciences, "The Digital Dilemma," Nov. 2007. Accessed: Jan. 26, 2024. [Online]. Available: https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma
2. Academy of Motion Picture Arts and Sciences, "The Digital Dilemma 2", 2012. Accessed: Jan. 26, 2024. [Online]. Available: https://www.oscars.org/science-technology/sci-tech-projects/digital-dilemma-2
3. International Federation of Film Archives (FIAF), "The Digital Statement." Accessed: Jan. 26, 2024. [Online]. Available: https://www.fiafnet.org/pages/E-Resources/Digital-Statement.html,
4. European Commission, "The Rolling plan for ICT standardization 2020." Accessed: Jan. 26, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/rolling-plan-ict-standardisation
5. EN17650:2022, "A framework for digital preservation of cinematographic works—The Cinema Preservation Package." [Online]. Available: https://standards.iteh.ai/catalog/standards/cen/cd05d0f8-88aa-46e7-957d-28b4e638cb68/en-17650-2022
6. CEN TR 17862:2022, "Guideline for the implementation of the Cinema Preservation Package (CPP) in EN 17650." [Online]. Available: https://standards.iteh.ai/catalog/standards/cen/151d64f0-54b0-43be-86cb-4a35cb89d3dd/cen-tr-17862-2022
7. METS, "Metadata Encoding & Transmission Standard." Accessed: Jan. 26, 2024. [Online]. Available: http://www.loc.gov/standards/mets/
8. METS Primer V1.6. Accessed: Jan. 26, 2024. [Online]. Available: https://www.loc.gov/standards/mets/METSPrimer.pdf
9. EBUCore, "Metadata Specifications." Accessed: Jan. 26, 2024. [Online]. Available: https://tech.ebu.ch/metadata/ebucore
10. PREMIS, "Preservation Metadata Maintenance Activity." Accessed: Jan. 26, 2024. [Online]. Available: https://www.loc.gov/standards/premis/
11. SMPTE, ST 2067-3:2020 "Interoperable Master Format — Composition Playlist," in ST 2067-3:2020, pp.1-35, 12 May 2020, doi: 10.5594/SMPTE.ST2067-3.2020.
12. DCI Digital Cinema Initiatives LLC, "Digital Cinema System Specification Version 1.4.4." Accessed: Jan. 26, 2024. [Online]. Available: https://www.dcimovies.com/specification/index.html
13. SMPTE, OV 2067-0-2021, "Overview Document - Interoperable Master Format," 2021.
14. CEN TC457 project team, "Open Source Software for CPP." Accessed: Jan. 26, 2024. [Online]. Available: https://gitlab.com/cen-pt457
15. International Organization for Standardization (ISO) 14721:2012, "Space data and information transfer systems—Open archival information system (OAIS)—Reference model," www.iso.org
16. EN15744, "Film identification—Minimum set of metadata for cinematographic works." Accessed: Jan. 26, 2024. [Online]. Available: http://filmstandards.org/fsc/index.php/EN_15744
17. Academy of Motion Picture Arts and Sciences, "Digital preservation forum." Accessed: Jan. 26, 2024. [Online]. Available: https://academydigitalpreservationforum.org/category/pillar-two/

About the Authors

Siegfried Fößel is heading the Moving Picture Technologies department at Fraunhofer IIS. In addition, he is responsible for the technology study program at the University for Television and Film HFF, Munich.

Heiko Sparenberg is Professor of Moving Image Technologies at RheinMain University of Applied Sciences. Before 2023, he worked as group manager for digital cinema at Fraunhofer IIS.

Nikolai Belevantsev graduated with a degree in computer science from FAU Erlangen-Nürnberg. At Fraunhofer, Belevantsev is working on the CreatiF project in collaboration with the HFF film school Munich, Germany, where he focuses on potential use cases of AI in film production.

Yi Lou oversees the archive project at the University of Television and Film, Munich, and holds a PhD in theater and film studies from the University of Munich.

Presented at the SMPTE 2023 Media Technology Summit. This paper is published here by kind permission of Faunhofer, IIS.

DOI: 10.5594/JMI.2024/NMRU3069
Date of publication: 1 April 2024

What's New for Members in 2024!

Access to the entire SMPTE Independent Study Virtual Course Library

Education is more accessible too. Members can access self-study courses at no additional cost, so you can learn at your own pace! In addition, instructor-led classes are available to members at a reduced price. Continue your education with SMPTE!

Access to Self-Study Video Library

EVOLVE with us
SMPTE

Join the Board of Editors

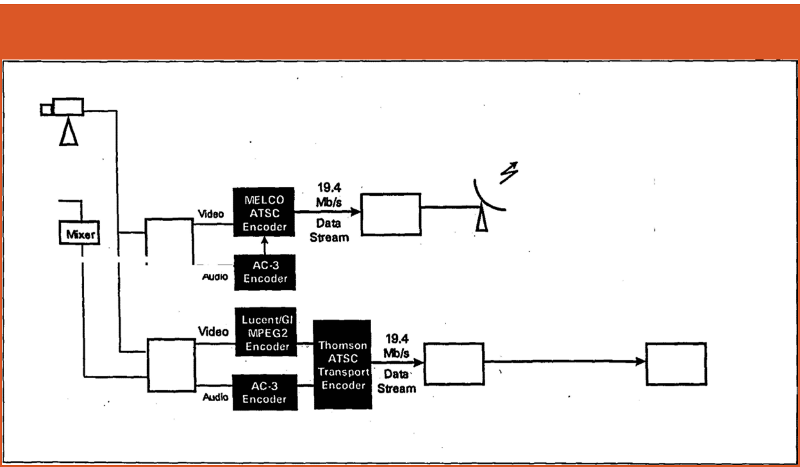
Volunteer to help shape and maintain the Journal's high editorial quality.

MI
MOTION IMAGING JOURNAL

SMPTE

RETROSPECTIVE

Historical Briefs From Past Issues



Origination, coding, satellite, and fiber distribution (Fig. 2 from *SMPTEJ*, April 1999, p. 203).

MICHAEL DOLAN

25 Years Ago in the Journal

The April 1999 *Journal* published in: “Design and Implementation of the ATSC Demonstration of HDTV at NAB’97” by Graham Jones: “In December 1996, the Federal Communications Commission issued the Rules and Order for the introduction of digital television in the U.S. To promote and publicize the introduction of high-definition digital television (HDTV), the Advanced Television Systems Committee (ATSC) committed to provide a demonstration of HDTV using the ATSC Digital Television Standard to be shown at the National Association of Broadcasters Convention, NAB’97, as part of the Special Technology Exhibits...Figure 2 shows the setup at WHD-TV.”

50 Years Ago in the Journal

The April 1974 *Journal* published in: “Report on Home Receiver Image Area Test” by R. J. Zavada: “The Geometry of the television picture tube has been evolving ever since the birth of the circular cathode-ray tube. These evolutionary changes, including size, have permitted set construction to be adapted to most home and commercial environmental needs. Our concern, however, is with those changes in tube face geometry which affect how much of the broadcast image is received by the home viewer. An early industry

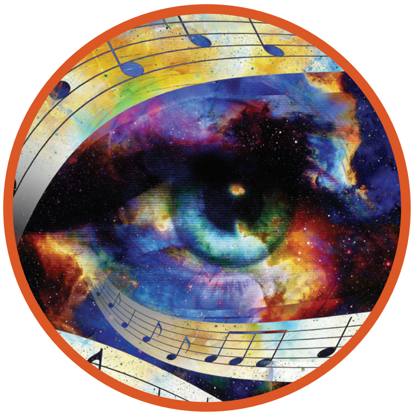
In this column we provide interesting historical briefs from the Journal articles of days past. The purpose of this column is primarily entertainment, but we hope it will also stimulate your thinking and reflection on the Society’s history, how far we have come in the industry, and (sometimes) how some things never change.

1545-0279/24©2024SMPTE

decision established the aspect ratio of the ideal television format at three units high and four units wide - this intentionally was exactly compatible with the 1.33.1 format used for early motion pictures and justly anticipated film as a significant source of program material. Once it was decided that the rectangular format was ideal, there immediately came the question of how to place the ideal format onto a tube face configuration that was at its inception circular. The amount of scan was one of the practical problems to be resolved; that is, should a complete rectangular picture be placed within the circular format, or should there be a significant amount of overscan such that the sides or the top and bottom were tangential with the circular tube face and that a certain amount of the transmitted or broadcast picture was lost to the home viewer?"

75 Years Ago in the Journal

The April 1949 *Journal* published in: "Possibilities of a Visible Music" by Ralph K. Potter: "Over 200 years ago a French mathematician and philosopher by the name of Louis Bertrand Castel proposed a visible music. He was probably the first to suggest specific possibilities of such a music and to attempt construction of an instrument. Castel thought of visible music as changing colored light and tried to associate color and musical tone. Others carried on the search in this direction, and until the late 1800's emphasis remained upon color. Experimental instruments built during this period were called "color organs." Then, following the color era, attention shifted to form...Where then must we look for a visible music? In particular, how would we know visible music if we were to see it? While this latter question seems, on first consideration, to be the kind we should prefer to leave to the philosophies, there is actually a simple answer. It is this: If we were to hear sound music and at the same time see a screen display that we feel is that music, the logical name for that display would be "visible music"! The conclusions of the analysis so far outlined arc that the possibilities of our having a visible music are excellent. The combination of such a visible music with the familiar audible type will offer the artist new opportunities for expression, and screen-and-sound audiences new and interesting entertainment."



"IF WE WERE TO HEAR SOUND MUSIC AND AT THE SAME TIME SEE **A SCREEN DISPLAY THAT WE FEEL IS THAT MUSIC,** THE LOGICAL NAME FOR THAT DISPLAY WOULD BE "VISIBLE MUSIC"!

100 Years Ago in the Journal

The May 1924 *Journal* published in: "Stereoscopy and Its Possibilities in Projection" by Hermann Kellner: "The words stereoscopy and stereoscopic are derived from the Greek words Stereo and Skopein which means "solid" and "to see." A stereoscopic picture is a picture that represents a solid aspect of an object similar to the impression gained when the object is looked at with both eyes in the natural way...Stereoscopic seeing can evidently be accomplished only when both eyes are functioning and when the left eye sees the left picture and the right eye sees the right picture...There is, however, a way of estimating distances with one eye which is of importance in motion picture projection. If we...assume a monocular (one-eye) observer placing his eye in succession in the positions of the right and left eye of the binocular observer, he will first see a picture like the one seen by the left eye of the binocular observer. When the observer moves toward ER the center post moves from its apparent position near the left rod towards the right post and it is by the amount of this shift in combination with displacement of the eye that he is able to form an opinion whether the center post lies in front of or behind the others and how far."

DOI: 10.5594/JMI.2024/XFBD7711
Date of publication: 1 April 2024

1545-0279/24©2024SMPTE

CO-LOCATED WITH **MPTS 2024**

MEDIA TECHNOLOGY CONFERENCE

15-16 MAY 2024 | OLYMPIA LONDON

THE TECHNOLOGY CONFERENCE FOR BROADCAST LEADERS

POWERED BY

THE MEDIA PRODUCTION & TECHNOLOGY SHOW

SMPTE UNITED KINGDOM

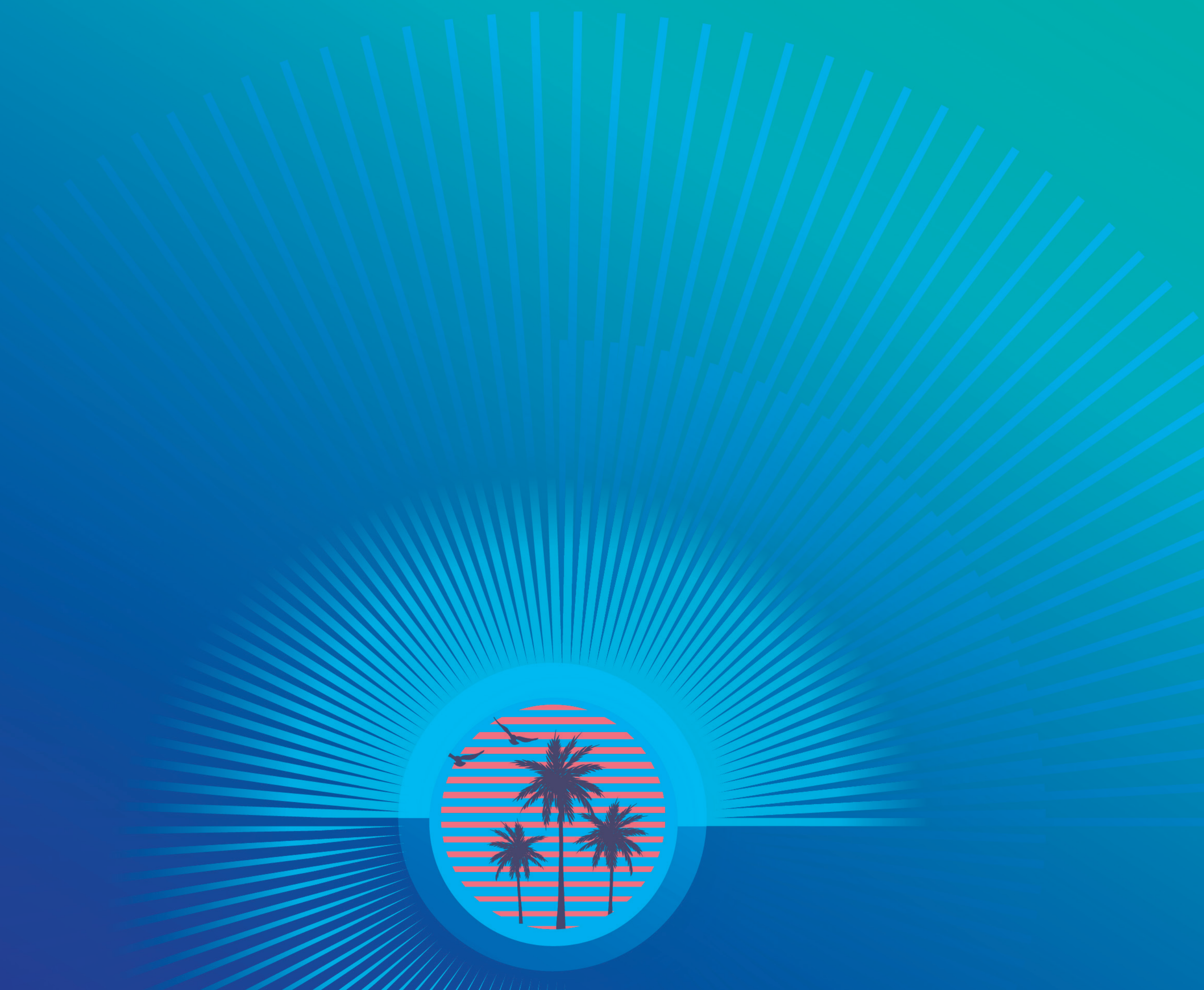
Topics include:

Emerging technologies | AI in media | IP workflows | Cloud for live production
Virtual production | Sustainability | Security | Skills | VR/XR

Register your interest today*

mediaproductionshow.com/technology-conference

* Spaces are limited and reserved for senior technology decision makers in the broadcast industry



SMPTE

MEDIA
TECHNOLOGY
SUMMIT

SAVE THE DATE

21-24 OCTOBER 2024