MMEDIA 2022

The Fourteenth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-942-3

April 24 - 28, 2022

Barcelona, Spain

MMEDIA 2022 Editors

Hiroshi Ishikawa, Tokyo Metropolitan University. Japan

MMEDIA 2022

Forward

The Fourteenth International Conference on Advances in Multimedia (MMEDIA 2022) continued a series of events aiming to provide an international forum by researchers, students, and professionals for presenting recent research results on advances in multimedia, mobile and ubiquitous multimedia and to bring together experts from both academia and industry for the exchange of ideas and discussion on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

The rapid growth of information on the Web, its ubiquity and pervasiveness, make the www the biggest repository. While the volume of information may be useful, it creates new challenges for information retrieval, identification, understanding, selection, etc. Investigating new forms of platforms, tools, principles offered by Semantic Web, opens another door to enable human programs, or agents to understand what records are about, and allows integration between domain-dependent and media-dependent knowledge. Multimedia information has always been part of the Semantic Web paradigm, but requires substantial effort to integrate both.

The new technological achievements in terms of speed and the quality of expanding and creating a vast variety of multimedia services like voice, email, short messages, Internet access, m-commerce, to mobile video conferencing, streaming video and audio.

Large and specialized databases together with these technological achievements have brought true mobile multimedia experiences to mobile customers. Multimedia implies adoption of new technologies and challenges to operators and infrastructure builders in terms of ensuring fast and reliable services for improving the quality of web information retrieval.

Huge amounts of multimedia data are increasingly available. The knowledge of spatial and/or temporal phenomena becomes critical for many applications, which requires techniques for the processing, analysis, search, mining, and management of multimedia data.

We take here the opportunity to warmly thank all the members of the MMEDIA 2022 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to MMEDIA 2022. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the MMEDIA 2022 organizing committee for their help in handling the logistics of this event.

MMEDIA 2022 Chairs

MMEDIA 2022 Steering Committee

Hiroshi Ishikawa, Tokyo Metropolitan University, Japan
José António Filipe, University Institute of Lisbon - School of Technology and Architecture, Portugal
Hanzhou Wu, Shanghai University, China
Max E. Vizcarra Melgar, University of Brasilia, Brazil

MMEDIA 2022 Publicity Chairs

Javier Rocher, Universitat Politècnica de València, Spain
Lorena Parra, Universitat Politècnica de València, Spain

MMEDIA 2022
Committee

MMEDIA 2022 Steering Committee

Hiroshi Ishikawa, Tokyo Metropolitan University, Japan
José António Filipe, University Institute of Lisbon - School of Technology and Architecture, Portugal
Hanzhou Wu, Shanghai University, China
Max E. Vizcarra Melgar, University of Brasilia, Brazil

MMEDIA 2022 Publicity Chairs

Javier Rocher, Universitat Politècnica de València, Spain
Lorena Parra, Universitat Politècnica de València, Spain

MMEDIA 2022 Technical Program Committee

Zahir Alsulaimawi, Oregon State University, USA
Pedro A. Amado Assunção, Instituto de Telecomunicações | Politécnico de Leiria, Portugal
Giuseppe Amato, CNR-ISTI, Italy
Hugo Barbosa, Lusofona University of Porto / Faculty of Engineering of the University of Porto, Portugal
Letizia Bollini, Free University of Bozen-Bolzano, Italy
Fernando Boronat Seguí, Universitat Politecnica de Valencia-Campus de Gandia, Spain
Dumitru Dan Burdescu, University of Craiova, Romania
Baoyang Chen, Central Academy of Fine Arts, Beijing, China
Trista Chen, Inventec Corporation, Taiwan
Jian-wei Liu, China University of Petroleum, Beijing, China
Minh-Son Dao, National Institute of Information and Communications Technology, Japan
Vincenzo De Angelis, University of Reggio Calabria, Italy
Franca Debole, Institute of Information Science and Technologies - Italian National Research Council (ISTI-CNR), Pisa, Italy
Jana Dittmann, Otto-von-Guericke-University Magdeburg, Germany
Vlastislav Dohnal, Masaryk University, Brno, Czech Republic
Filiz Ersoz, Karabük University, Turkey
Taner Ersoz, Karabük University, Turkey
Manuel Alberto M. Ferreira, University Institute of Lisbon - School of Technology and Architecture, Portugal
José António Filipe, University Institute of Lisbon - School of Technology and Architecture, Portugal
Tolga Genc, Marmara University, Turkey
Konstantinos Gkountakos, CERTH (Centre For Research & Technology Hellas) | ITI (Institute of Informatics), Greece
Jun-Won Ho, Seoul Women's University, South Korea
Yin-Fu Huang, National Yunlin University of Science and Technology, Taiwan
Hiroshi Ishikawa, Tokyo Metropolitan University, Japan
Salma Jamoussi, University of Sfax | Higher Institute of Computer Science and Multimedia, Tunisia
Dimitris Kanellopoulos, University of Patras, Greece
Sokratis K. Katsikas, Norwegian University of Science and Technology, Norway
Panos Kudumakis, Queen Mary University of London, UK
Fons Kuijk, Distributed and Interactive Systems - CWI, Amsterdam, Netherlands
Cristian Lai, CRS4 - Center for Advanced Studies, Research and Development in Sardinia, Italy

Copyright Information

Table of Contents

Comparison of Two Approaches for Human Tense Situation Analysis in Car Cabin

Quentin Portes
Renault Software Lab
Toulouse, France
quentin.q.portes@renault.com

José Mendes-Carvalho
Renault Software Lab
Toulouse, France
jose.mendes-carvalho@renault.com

Julien Pinquier
IRIT, Paul Sabatier University, CNRS
Toulouse, France
julien.pinquier@irit.fr

Frédéric Lerasle
LAAS-CNRS, Paul Sabatier University
Toulouse, France
lerasle@laas.fr

***Abstract*—The use of audio, video and text modalities to simultaneously analyze human interactions is a recent trend in the field of deep learning. The multimodality tends to create computationally expensive models. Our in-vehicle specific context requires recording a database to validate our approach. Twenty-two participants playing three different scenarios ("curious", "argued refusal" and "not argued refusal") of interactions between a driver and a passenger were recorded. We propose two different models to identify tense situations in a car cabin. One is based on an end-to-end approach and the other one is a hybrid model using handcrafted features for audio and video modalities. We obtain similar results (around 81% of balanced accuracy) with the two architectures but we highlight their complementary. We also provide details regarding the benefits of combining different sensor channels.**

***Keywords*—Human interactions, multimodality, data fusion, audio & video features, end-to-end.**

I. Introduction

Today, most of the data available on the Internet is saved under a video or an audio format. Sight and hearing are the main channels used by the brain to understand and decode human interactions. Voice is implicitly processed into words by the brain. In case of multiple speakers, improving the process of interaction analysis could lead to increase performances of sentiment, emotion and dialog analysis. These multiple speaker situations are common in the industrial context, *i.e.* the necessity to improve social media filtering, human-machine interaction understanding, brand monitoring, etc. In the automotive context, it will answer safety concerns (*i.e.* taunting, bullying or, in the worst case, aggression) linked to the new usages of cars (*i.e.* socializing, vehicle sharing, autonomous cars, etc.). Our aim is to detect the signals leading to these situations in order to anticipate and avoid them.

To address this issue, we can analyze the passengers' interactions thanks to cameras and microphones on boarded in the car cabin. These two sensors generate three modalities (video, audio and the text transcribed from the audio), which can be combined to significantly improve the performances of human tense situation predictions.

Today, these modalities are usually analyzed with deep learning approaches. We use Bidirectional Encoder Representations from Transformers (BERT) architecture [1] (English language), Roberta and CamemBERT models [2] (French language) for text analysis. They have improved the global performance in question answering, text summarizing tasks, etc. Recent work uses the transformer model for text dialog analysis [3] [4].

For the video modality, 2D and 3D [5] [6] convolutional approaches are the predominant architectures to analyze images and video.

The most common technique regarding audio analysis is the extraction of audio features over a short sliding window with a framework, such as open SMILE [7]. Then, they are usually fed to a sequential model like Long Short-Terme Memory (LSTM) [8].

One way to improve performances of such models is to combine the audio, video and text analysis. This approach contains more information than the video and audio modalities separately [9]–[13].

The automotive context is an embedded system with some associated constraints: execution time, limited computational resources, memory access, etc. Processing and analyzing three modalities with deep learning algorithms tend to induce large models. To deal with the multimodality and the embedded constraints, one solution is to design a hybrid model with one compact model based on handcrafted features running on embedded hardware and one larger model with an end-to-end architecture running on a cloud platform.

The four challenges identified are the following:

- The availability of a public in situ dataset.
- The fusion between non-heterogeneous modalities like video, audio, and text.
- The complexity to model human interactions.
- The embedded constraints.

Actually, to the best of our knowledge, the literature does not deal with all of these issues at the same time. We are addressing them hereafter. This research focuses on recording an exploitable dataset for industrial applications and then designing two different approaches showing the benefits of the multimodality for detecting tense situations in the car's cabin.

We differ from the literature by our realistic in-situ dataset and our two complementary multimodal models. We also present two different strategies of late fusion.

Section II introduces a literature review on multimodal dialog analysis. In Section III, we detail the protocol used to record our own dataset and its specifications. Section IV provides details and compares our two multimodal approaches for the classification of tense human interactions. Finally, Section V present our results.

II. Related work

The modern dialog, interaction and conversation analyzing models are based on text [14], [15]. Recent investigations, with new approaches such as multimodality, show the benefits of exploiting information from different channels. Every multimodal model on sentiment analysis fields outperforms unimodal architecture ones [9], [16]. Due to the heterogeneity of the modalities (audio, video and text) used in these architectures, the features are extracted per modality. Then, a final, more or less complex, late fusion is applied to obtain better results. The end-to-end models extracting the features tend to be computationally expensive compared to handcrafted approaches. They also need more data to be trained. Most of the time, handcrafted and end-to-end models are compared only on prediction performances. In the context of human behavior understanding, we can capitalize on the full potential of both techniques. Indeed, the study of human interactions, sentiment analysis or emotions represents some knowledge that we can directly inject in a model. Conversely, we can automatically let end-to-end models find features. These two opposite techniques can be complementary in some scenarii.

Our preliminary works are based on a public dataset like MOSI [12] [17]. We identified work on multimodal conversation analysis such as [18] [19] that train on this previous dataset. Additionally, they are only focusing on sentiment and emotion conversation analysis.

Hierarchical Attention Network (HAN) architecture [20] is performing very well as the Transformer [1] on document analyzing. Recent approaches, such as [3], are using Transformer for dialog analysis. As we are working on oral text and a small dataset, the HAN approach seems to be the most appropriated.

Regarding interaction analysis, the speaker's previous behaviors are crucial to hold. Nowadays, the deep learning architecture is not able to process extensive videos. The use of stateful temporal models [8] in our approach will allow us to keep track of the information over scenario duration.

The investigations concerning the car cabin passenger interactions are very scarce and remain a scientific challenge.

III. MULTIMODAL DIALOG CORPUS IN VEHICLE

In this section, we detail the protocol used to record our multimodal dataset. The aim is to classify three different types of interactions. The first one is the ”normal/curious” category where two participants have a cordial discussion. The second one is the ”argued refusal”, where the rear passenger refuses cordially the driver's proposition. The last one is a full refusal of the driver proposition, called: ”not argued refusal”. The insistent seller scenario has been chosen instead of an aggression scenario for two reasons. The first one is our objective to find discussion resulting in aggression and not physical aggression. The second is due to protocol reliefs reasons. Indeed, willing to play realistic aggression scenarios, obliging to follow a psychological protocol setup for the different subjects would be very restrictive.

A. Purpose of the dataset

We recorded the interactions between two passengers in a car's cabin (see Fig. 1). One driver and one rear seat passenger (right side) are playing predefined non-scripted scenarios. Subjects are French volunteers without any acting skills.

Each pair of participants is recorded for 7 minutes, scheduled in a session of four continuous stages. This paper only focuses on the acting stage:

1) 60s of silence,
2) **180s of acting**,
3) 60s of silence,
4) 120s of interaction with the In-Vehicle Infotainment (IVI).

During the acting stage, the driver always plays the same role of an insistent seller and the passenger plays one of the three following behaviors:

- ”be curious about the driver proposition”,
- ”refuse the proposition with argumentation”,
- ”refuse categorically the proposition”.

We set up a double-blinded scenario. The driver and the passengers never knew the situation that has to be played beforehand. In this configuration, we can say that the driver undergoes the situation.

B. Acquisition setup

We equip a Dacia Duster with six cameras, four microphones and one screen placed on the hood of the car. The screen is in front of the driver view and also visible by the passenger. Its use is motivated to indicate when the subjects have to change the acting phase and stream a video of the road to captivate the driver's attention due to the stationary car. The interactions with the car are available (wheel, gear lever, etc.).

1) Video steaming: All the cameras present in the setup have different resolutions, angles of view and lenses. Our approach privileges the camera #2 because it has the best view and lighting quality. It is a manual-focus camera of recording resolution 1920×1080 pixels. It is placed in order to have a front angle of view, see Fig. 1.

The other cameras will be considered for future investigations.

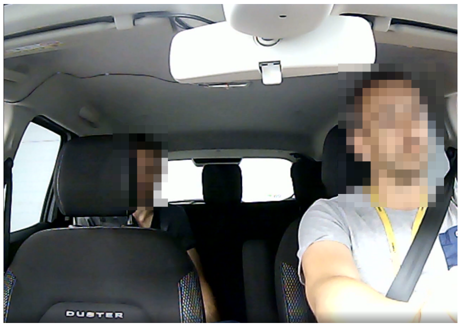


Fig. 1. Field-of-view of the camera #2.

2) Audio streaming: Four identical microphones Brüel&Kjaer prepolarized 1/4 inch Type 4958 are set in different vehicle areas recording the audio stream. Our approach only uses the ceiling driver's microphone because it is the only area used by the car manufacturer.

We recorded all the video and audio streams in RAW format (no live compression) in the concern of not using too much computational power.

C. Preprocessing and annotation of the dataset

Once recorded, a step of post-processing is mandatory. Indeed, the recording process generates a temporal delay between video and audio streams. The six videos and four audio streams are synchronized with Adobe premiere pro. Ultimately, the videos are compressed in MPEG-4 format.

The third modality (text) is obtained by transcribing the audio stream. After some experiments, we avoid the automatic speech transcription (ASR) such as *Amazon transcribe* or *Google speech to text* due to their very high word error rate. In our oral context, repetitions, interjections and isolated words are the most important parts of the dialog. Furthermore, the sentences are potentially poorly constructed (subject-verb-complement). ASR techniques are inefficient in that case.

We use the ELAN software to transcribe the dataset. It is a manual annotation tool for video and audio data. The audio stream is transcribed into utterances for each actor, resulting in a total number of 2026 utterances. An utterance is a continuous unit of speech beginning and ending with an explicit pause. The transcript is reviewed by a peer.

To reduce the annotation time, we annotate at the scenario level in comparison to other datasets [17], where the annotations are at the utterance level. An entire recording sequence is annotated at the beginning of the recording. This choice has some repercussions: for instance, it induces wrong labels if the subjects play their roles not in adequacy with the asked one. We will come back to these issues later in the qualitative analysis (see Section V-B).

D. Specifications and understanding of the corpus

Finally, the dataset contains 44 videos for 22 participants (4 women/18 men). Each pair of participant plays once as a driver and once as a passenger in a random order. The accumulated interactions give on average 46 sentences per video, for a total of 2026 sentences. This represents 21 966 words containing 2082 unique words. We get 54 min for the ”curious” class, 27 min for the ”argued refusal” class, and 27 min for the ”not argued refusal” class, which represent a total of 1h48. An asymmetry in the amount of data recorded is added to take into consideration the fact that in real situations the curious class would be the usual behavior.

By examining the video, we notice that the video modality is less informative than the audio and text ones. Indeed, the passengers are mostly static due to the car context and the belt as well as the driving task restricting the movements. We also detect this outcome in sentiment or dialog analysis based on multimodal datasets [16] [21].

The analysis of the dataset over time shows patterns in the drivers' and passengers' behaviors. Humans are not swapping their emotions or behaviors at a high frequency. Taking this information into account, we decide to plot the features as a function of time for a 15s analyzing window; values higher than 30s result in flat curves with no possible deduction. This Github link[1] makes available the plotted chart. The local descriptor plots are inspired from [22].

After examining the audio-video streams and analyzing the charts, we are able to focus on seven hand-crafted features, as indicated below. Four of them are generated by ”the mean talking” and ”mean duration” for the two passengers and the three remaining are the ”mean silence”, the ”eye contact” and the ”passenger visibility”:

- Mean talking - In a normal conversation, the average talking tends to be equitably distributed among the participants.
- Mean duration - It is the average duration of the utterances. Complementary to the mean talking, the length of a speech is a good indicator of who is dominating the conversation and who wants to close the dialog.
- Mean silence - The mean silence is an indicator of the intensity of a dialog. The more silence there is, the more the discussion is poor and tends to be in the refusal situation.
- Eye contact - It is the frequency at which the driver is looking into the interior rear-view mirror. Eye contact is a natural behavior when talking to someone. As the driver is focused on the road and on the driving task, he has no other choice but to look at the rear-view mirror to see its interlocutor.
- Passenger visibility - It is the frequency at which the passenger is seen by the camera. It is a good indicator of the passenger's interest in the conversation. We naturally reduce the distance with our interlocutor when we are engaged in a discussion. In the car discussion context, the rear passenger can move forward between the two front seats. On the video stream, it results in seeing (or not) the rear passenger.

For the text modality, we calculate the frequency distribution of words and the term frequency-inverse document frequency

[1] https://github.com/QuentinPrts/MMEDIA_2021

(TF-IDF) [23] to find if there are specific distributions of words associated with a given scenario. These approaches are very common in text mining and analyzing. The TF-IDF delta between the two opposite classes (”curious” and ”not argued refusal”) exhibit the 10 following most important delta words: *je (I), pas (not), vous (you, second-person plural), ouais (ok), tu (you, second-person singular), non (no), moi (me), oui (yes), donc (so)* and the *ah* interjection. The text modality is not rich (as a reminder, we have 2082 different words).

In the chart, we observe two transition phases. The first one is the setting up: the subjects could not be insistent or categorical in their refusal to lead a ”bad acting” in the first 30s of each scenario. The second is at the end: subjects run out of inspiration, causing shortness of breath for the last 20s of each scenario. This changeover is due to the individuals playing the scenarios: they are volunteers and not real actors.

IV. MULTIMODAL ANALYSIS

After analyzing the dataset, we implement two different approaches, one end-to-end model (noted E2E) and one based on handcrafted features (noted H). They have to process data to classify the input stream into three classes corresponding to our three scenarios (”curious”, ”argued refusal”, ”not argued refusal”). The two architectures are detailed in the following sections. Fig. 2 illustrates our two approaches. First, we implement a dedicated model for each modality and evaluate their performances after fusing their outputs. Then, the modalities are converted into a generator of features for a multimodal fusion purpose.

As the basic analysis of the text modality is not performing very well, we decided to implement a deep learning model. It will be used for both approaches presented in the following section.

A. Text analysis

We face a major problem in the text modality. Indeed, every framework and pre-trained models such as Spacy [24], NLTK [25], BERT [1] are well suited for English analysis but perform very badly on the French language. The existing French alternatives are very limited because they are based on old or written French. Thus, we did not obtain sufficient results on the transformers model named Camen-BERT [26] which is trained on Wikipedia text. The poverty of our text makes the basic approach (TF-IDF and embedding + LSTM model) inefficient.

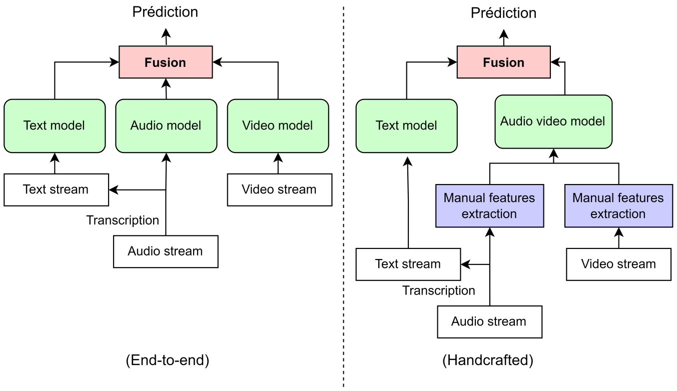


Fig. 2. The two approaches implemented: E2E (left) and H (right).

Ultimately, we implement the Hierarchical Attention Network (HAN) [20], which was originally designed for text document classifiers. We choose this architecture because it has the ability to focus on both word and sentence levels thanks to its attention mechanism.

We modify the original implementation by replacing the basic Gated Recurrent Unit (GRU) layer of the sentence level by a stateful GRU. This modification allows the model to keep track of the hidden state over time, improving the global performances.

The hyper-parameters of this model are tuned empirically. The input of the embedding is of size 500 which is the number of words the most represented in the dataset, and the output is of size 100. Each one of the two GRUs has 16 cells.

B. Handcrafted approach

This first approach consists of combining text and high-level audio-video hand-crafted features. We extract a total of 32 features with the text model and four features from the seven aforementioned raw handcrafted features.

1) Audio-video analysis: Among the seven features, two are extracted from the video stream. The first one, named ”eye contact”, is calculated using the extracted face with Dlib [27] and openCV [28] then hyperface [29] to generate the Euler angles of the head. This process is applied to each frame of the dataset. Finally, a K-means clustering algorithm on the Yaw and Pitch axis determines the couple of Euler angles when the driver is looking in the rear-view mirror. The tilt axis does not provide additional information in the car context.

For the ”passenger visibility”, we use Dlib and openCV to detect the face of the rear passenger on each frame. It is a binary feature, set at 1 if we detect the face of the rear passenger, 0 otherwise.

The five remaining features are the ones detailed in Section III-D: ”the mean talking” and ”the mean duration” for the driver and the passenger, and the ”mean silence” which is common to both of them.

Finally, these seven features fed a Multi-Layer Perceptron (MLP). It is designed with two hidden layers of four neurons each and one output layer generating the prediction.

2) Temporal fusion of our cues: Adding a temporal late fusion is necessary in our case because the stateful HAN is not sufficient to capture all the temporal information. The Perceptron model has no ability to capture temporality in the data. Furthermore, a late fusion is the usual strategy in case of non-heterogeneous modalities.

The fusion concatenates all features extracted from the three modalities (see Fig. 3). The unimodal models are modified to extract 32 features from the text and four from the audio-video model. It results, after concatenation, in a vector of size 36. Then, they are stacked for each analysing window of 35s to

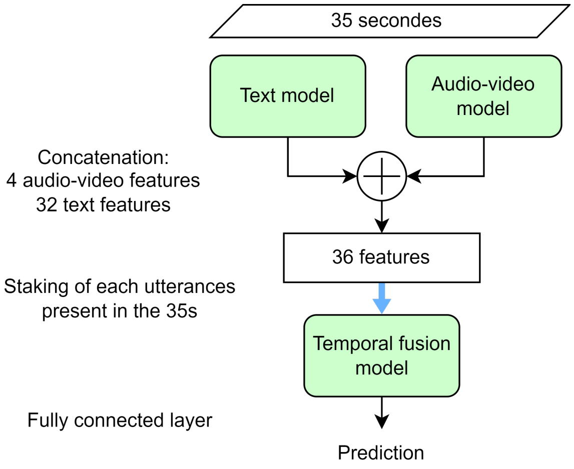


Fig. 3. The temporal fusion for the model H.

finally feed a stack of two stateful temporal Recurrent Neural Network (RNN) named GRU. See [8] for a complete review of the RNN. Finally, a Fully Connected (FC) layer predicts the label. The concept of stateful model is detailed in the Section IV-D.

C. End-to-end approach

This section details the end-to-end approach where the raw data are given in input of the model to directly make the prediction.

1) Audio analysis: In this new approach, we use OpenSMILE [7] with the configuration file emobase2010 [30] to automatically extract the features. We only use the audio stream of the passenger for more than one second. This process filters the interjections like: 'euh', 'ok', 'okay' and the repetitions of words very common in oral language. They are filtered because they do not increase the performances and can even degrade them. Moreover, the stress on the word is very weak in our corpus and more generally in the French language compared to the English language. The stress is more noticeable at an utterance level rather than at the word one in the French language [31]. OpenSMILE allows us to calculate the average value over a period of time. As a result, we have 1581 features per utterance. Finally, these features are stacked for each utterance in the window analysis and sent sequentially to a stateful recurrent network of two layers of 12 cells GRU. Then, a fully connected layer makes the final prediction. The matrix feeding the model is the size of the number of utterances in the 35s of the analysis window by the 1581 features.

2) Video analysis: Recall that the video modality is the least informative feature in our context and is also the least informative in the sentiment and emotion analysis literature [9] [11] [16]. In fact, only the head, the eyelid and the mouth movements give us information. This information is also limited because the driver must not deviate from his driving simulation task. As a reminder, the simulation task is a driving video in the first person view shown on the screen placed on the hood of the car. From our experience acquired with the MOSI corpus use [12], we first experiment the R3D approach [6]. The results were not conclusive. We then implement several other models with the ability to model the temporality:

- convLSTM [32],
- 2D Convolutional Neural Network (CNN) + LSTM,
- R3D + LSTM,
- optical flow [33] + R3D,
- optical flow [33] + 2D CNN + RNN.

All these architectures cited did not give satisfactory results. The models are able to converge during the training phase, but the results collapse during the validation phase. There are several hypotheses explaining this phenomenon: maybe an overfitting problem, an insufficient amount of data, or maybe the models are not able to catch the right features for the classification task.

These results lead us to test two last solutions. The first one uses a vector of 128 features to encode the driver's face in each image. We use the Dlib library [27] to extract them. Then, the features are stacked in 35s windows and sent to a GRU or LSTM model. The results are still not convincing.

The second method extracts the key points of the face. The principle is to retrieve 68 facial landmarks (the contour of the eyes, the face, the nose, etc.) defined by its image coordinates. They are computed for each frame using the Opencv and Dlib libraries. This approach gives good results compared to all the other implementations. We add the head orientation angles as described in Section III-D to improve the performance. Finally, a total of 142 facial features including the three head angles encode the driver's face.

Once the data are processed, they feed a neural network of two layers stateful GRU of four cells each. The matrix feeding the model is the size of the number of frames in the 35s of the analysis window by the 142 features.

3) Fully connected fusion of our cues: In this approach, the temporality is taken into account at the modality level (with the use of stateful GRU) in contrast to the handcrafted approach where it is at the fusion level. We modify the unimodal model to extract features. As each different modality does not have the same impact on the prediction performance, we empirically determine the number of features per modality to obtain the best performances. A total of 10 features are concatenated, four regarding text and audio features and 2 for the video one. Then, a fully connected (FC) layer built of 30 parameters makes the final prediction. See Fig. 4 for a representation of this approach.

D. Implementation Details

Setting the free parameters of the architecture and the training process are really important to deal with the multimodality and temporal context.

Foremost, we empirically set the sliding analyzing window to $T = 35s$.

During a dialog, the situation can evolve and catching this gradation gives a lot of information. As long as videos must be processed by algorithms with smaller analyzing windows, it is

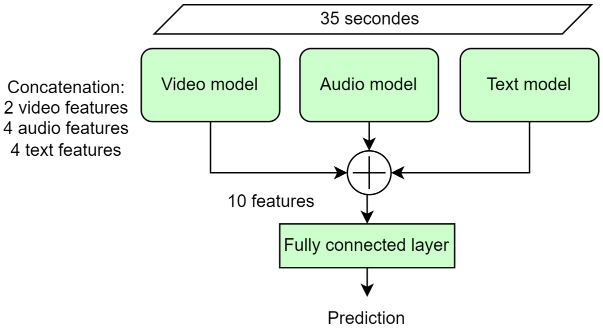


Fig. 4. The fully connected fusion for the model E2E.

important to keep track of the context between each analyzing window. We implement this concept by using stateful GRU. RNN only remembers what happened within a sequence. A sequence can be a set of sentences, a set of features, etc. At the initial time point of every passed sequence, the hidden states are initialized at 0, which means that the previous information is lost. In our approach, we initialize at each iteration the hidden state with the one generated at the previous analyzing window. This process keeps track of the evolution of all the features from the beginning to the end of the video.

Stateful RNN must be trained video by video. Each video is cut on the fly into approximately $180/35 = 5$ subsequence video clips. Then, they are fed chronologically one by one to the model. This training approach generates only $44 * 5 = 220$ training samples. In order to increase the training set, we shift the beginning of the analysis window to generate 400 samples. This consists in passing multiple times over each video. At each iteration, the starting point of the analyzing window is shifted.

Recall that the limit of our dataset forces us to discard the first 30s of our training samples. We delete on the fly these files during the training and validation phases.

The last method used to train the multimodal model is the pre-training techniques. All unimodal models are firstly trained to reach their best accuracy point and be saved. Then, at the beginning of the multimodal training phase, each previously saved model is loaded to initialize the multimodal one. This method is mandatory in our approach, otherwise the multimodal model is not able to converge. Freezing the weight of the loaded model (except for the fusion model) is considered, but it leads to poorer performance results.

On a multi-class problem, we use the cross entropy loss defined as in equation (1).

$$\operatorname{loss}(\hat{y}, \text{ class }) = -\log \left(\frac{\exp(\hat{y}[\text{ class }])}{\sum_i \exp(\hat{y}[i])} \right) \tag{1}$$

where $\hat{y}$ is the output score of the model for the corresponding class.

V. Evaluations and associated analysis

In this section, we present the quantitative evaluations for both approaches and a qualitative analysis.

A. Quantitative Evaluations

When we work on behavior or emotion analysis, the speaker dependency is a key point. The idea is to evaluate the abilities of the algorithm to generalize when it deals with a new speaker. For this purpose, we generate five different cross-validation sets by selecting 80% of the speakers for the training phase and 20% for the validation phase. More specifically, there are in the train set: 36 videos representing 1620 utterances generated by 18 speakers and for the validation set: eight videos totaling 405 utterances generated by four speakers.

The balanced accuracy is defined in equation (2). It is mandatory when we do not have a balanced number of samples in each class.

$$\text{balanced-accuracy } (y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1 \left(\hat{y}_i = y_i \right) \hat{w}_i \tag{2}$$

It is the macro-average of recalled scores per class i with associated weight $\hat{w}_i$ relative to the inverse prevalence of its true class y_i. The $\hat{y}_i$ is the inferred value of the sample i.

We obtain the following results (see Table I). This is the mean of the five cross-validation sets.

The results obtained with these two approaches are quite similar. Indeed, we obtain 81.6% of balanced accuracy with the end-to-end model and 81% with the handcrafted approach. The approach using the handcrafted (H) features is more consistent with a standard deviation below the end-to-end approach (E2E). The (H) architecture does not contain enough parameters on the video or audio alone to allow a classification by modality. The standard deviation of (E2E) is likely due to the audio and video modalities which are difficult to exploit. As a reminder, we were able to obtain convincing results by using only the rear passenger data for audio and the driver's face for video. Additionally, the literature [34] shows the existence of a threshold in the quantity of data for which the end-to-end approaches outperform classical approaches (statistical, machine learning, etc.). Below this threshold, the classical techniques obtain the same or better performances as the end-to-end one. Our results and the size of our corpus seem to indicate that we are in this situation. Increasing the amount of the data could therefore solve the issue.

The case of speaker dependency in the selection of train/test data is the least favorable for a neural network and not representative of real-world applications. Indeed, in the case

TABLE I
Results and comparison of the two approaches. SD refers to Speaker Dependant.

Model	Modalities	Balanced accuracy
End-to-end (E2E)	Video	65.6% ± 4
	Audio	70.6% ± 4,9
	Text	70% ± 0.8
	Audio + video	61% ± 3.9
	Video + Audio + Text	**81,6% ± 5.9**
	Video + Audio + Text SD	***88.2%***
Handcrafted (H)	Text	70% ± 0.8
	Audio + Video	60% ± 1.12
	Video + Audio + Text	**81% ± 1.2**

of a smartphone assistant, a ”world” model is specified to a new user with a new training phase based on some samples of his voice. For example, on a new Android device, when the user starts to use Google assistant, it asks the user to repeat a few times ”Ok Google”. To reproduce this configuration, we train our end-to-end model on the first 90s of each video and test it on the remaining 90s. This approach gives a balanced accuracy of 88.2%. It shows the benefit of a specialization phase and partially shows our shortage of data.

B. Quantitative evaluations and Juxtaposition of the two approaches

After analyzing the miss-classified files, we observe some issues leading to these miss-classification. A few participants did not play their role in adequacy with the asked scenario or they took a very long time to engage in the discussion. Two specific issues also lead to bad results: (i) on one video the voices of the two speakers are very low compared to the other’s recording; (ii) on another video, the driver has a very bad posture resulting in a half-visible face.

Typically, in literature, the performances of the different approaches are compared. We argue that the two presented approaches are complementary. If we examine the wrong/right classifications of each model for the same test file, we notice that errors are not made on the same analysis window. These two models can be complementary in their decision-making.

Fig. 5 shows the confusion matrices for our two models on the same cross-validation file. The end-to-end approach has a better ability to classify the videos of the ”curious” and ”categorical refusal” classes which are the most opposed classes. The handcrafted based model performs well on the ”argued refusal” class.

In our application context, one hybrid solution could be to use the embedded model, *i.e.* Handcrafted (H), to establish a first diagnosis of the situation and then send the video data to a cloud platform to inference with the end-to-end model. This choice lead to reduce the cost of data transmission to a cloud platform.

	Predicted classes					
	Model (E2E)			Model (H)		
Real classes	cur	ref_arg	ref_cat	cur	ref_arg	ref_cat
cur	13	0	0	10	3	0
ref_arg	1	4	2	0	7	0
ref_cat	2	2	15	1	9	9

Fig. 5. Comparison of the two confusion matrix for a same cross-validation set. (H) refers to the handcrafted model and (E2E) to the end-to-end one. cur denotes the ”curious” class, ref_arg describe the ”argued refusal” and ref_cat stands for the ”not argued refusal” class.

VI. Conclusion

This paper compares two multimodal approaches for the analysis of interaction in a real vehicle context. The performances obtained with these models are promising with 81% of balanced accuracy for the handcrafted model and 81.6% for the end-to-end approach. We also show the benefits of the fusion of different modalities and the complementary of our two approaches.

The embeddability capability of neural networks and the real application context is often omitted in the literature and even more in multimodal systems. Our future work will focus on integrating the two detailed architectures on a specific automotive platform to evaluate the embedding performances.

Acknowledgment

This work has been carried out under the funding of an industrial doctorates fellowship from the National Association for Research and Technology (ANRT), France.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[2] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot, “CamemBERT: a Tasty French Language Model,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219.

[3] B. Santra, P. Anusha, and P. Goyal, “Hierarchical Transformer for Task Oriented Dialog Systems,” *arXiv:2011.08067 [cs]*, Mar. 2021.

[4] D. Chen, H. Chen, Y. Yang, A. Lin, and Z. Yu, “Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems,” *arXiv:2104.00783 [cs]*, Apr. 2021.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *The IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015.

[6] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3d residual networks for action recognition,” *arXiv:1708.07632 [cs]*, 2017.

[7] F. Eyben, M. Wöllmer, and B. Schuller, “opensmile – the munich versatile and fast open-source audio feature extractor,” in *ACM Multimedia*, 01 2010, pp. 1459–1462.

[8] Y. Yu, X. Si, C. Hu, and J. Zhang, “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures,” *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[9] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 873–883. [Online]. Available: http://aclweb.org/anthology/P17-1081

[10] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, “An ensemble approach to utterance level multimodal sentiment analysis,” in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 145–150.

[11] A. Agarwal, A. Yadav, and D. K. Vishwakarma, “Multimodal sentiment analysis via rnn variants,” in *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science Engineering (BCD)*, 2019, pp. 19–23.

[12] Q. Portes., J. Carvalho., J. Pinquier., and F. Lerasle., “Multimodal neural network for sentiment analysis in embedded systems,” in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, INSTICC. SciTePress, 2021, pp. 387–398.

[13] Q. Portes, J. Pinquier, F. Lerasle, and J. M. Carvalho, “Multimodal human interaction analysis in vehicle cockpit,” in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2118–2124.

[14] R. Li, C. Lin, M. Collinson, X. Li, and G. Chen, “A dual-attention hierarchical recurrent neural network for dialogue act classification,” in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 383–392.

[15] Y. Luan, Y. Ji, and M. Ostendorf, “Lstm based conversation models,” 2016.

[16] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. B. V. Subramaanyam, “Benchmarking multimodal sentiment analysis,” *arXiv:1707.09538 [cs]*, 2017.

[17] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” 2016.

[18] W. Li, W. Shao, S. Ji, and E. Cambria, “BiERU: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis,” *arXiv:2006.00492 [cs]*, Feb. 2021.

[19] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive rnn for emotion detection in conversations,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, Jul. 2019.

[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489.

[21] Q. Portes, J. Carvalho, J. Pinquier, and F. Lerasle, “Multimodal neural network for sentiment analysis in embedded systems,” in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, INSTICC. SciTePress, 2021, pp. 387–398.

[22] B. Bigot, J. Pinquier, I. Ferrané, and R. André-Obrecht, “Looking for relevant features for speaker role recognition (regular paper),” in *INTERSPEECH, Makuhari, Japan, 26/09/10-30/09/10*. http://www.isca-speech.org/: International Speech Communication Association (ISCA), 2010, pp. 1057–1060.

[23] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[24] M. Honnibal and M. Johnson, “An improved non-monotonic transition system for dependency parsing,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1373–1378. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1162

[25] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP ’02. Association for Computational Linguistics, 2002, p. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[26] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, “CamemBERT: a tasty French language model,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 7203–7219.

[27] D. E. King, “Dlib-ml: A machine learning toolkit,” *J. Mach. Learn. Res.*, vol. 10, p. 1755–1758, dec 2009.

[28] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[29] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[30] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” *INTERSPEECH 2010*, p. 4, 2010.

[31] J. Vaissière, “Cross-linguistic prosodic transcription: French vs. English,” in *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, N. Volskaya, N. Svetozarova, and P. Skrelin, Eds. St Petersburg State University Press, 2002, pp. 147–164.

[32] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. kin Wong, and W. chun Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” 2015.

[33] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2017/IMKDB17

[34] M. Z. Alom, T. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. Nasrin, M. Hasan, B. Essen, A. Awwal, and V. Asari, “A state-of-the-art survey on deep learning theory and architectures,” *Electronics*, 03 2019.

Epipolar Shift Compensated Light Field Video Quality Metric

Nusrat Mehajabin, Dan Jin, Rui Yao, Thomas Dykstra, Mahsa Pourazad, Panos Nasiopoulos
Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, Canada
e-mail: {nusratm, pourazad, panosn}@ece.ubc.ca,{dysjin, rruiyao, dykstr01}@student.ubc.ca

***Abstract*—Traditional video quality metrics are unsuitable for the Light Field (LF) video content as these metrics do not account for the structural and angular relationships among the various viewpoints found in LF content. While there is a growing amount of light field video content being produced for increasing application demand, there is currently no standardized objective method for measuring the quality of these videos. In this paper, we propose an objective quality metric for evaluating the spatial and angular quality of light field video content. We achieve this goal by leveraging the Epipolar Plane Images (EPI) along the horizontal, vertical, and diagonal views, on which we perform statistical analysis to determine the quality of the LF content. We also present our results and discuss our findings and future work on this topic.**

Keywords -Light Field; SSIM; PSNR; Objective Quality Metric; Epipolar Plane Image.

I. INTRODUCTION

Light field video is an interesting new technology that holds great promise. By recording video using multiple cameras [1] that are all pointing at the same scene, one can perform operations, such as changing perspective, “peeking” around objects in the foreground to see some of the background [2], and changing image focus in post-production [3], etc. Generating, transmitting, and rendering Light Field (LF) content is a growing field of research [4]. To meet different application demands, the industry must compress [5], synthesize, calibrate [6], and perform other operations on the original content. Therefore, there is a need for a quality metric to make sure that the processed content preserves the original spatial and angular relationships.

Prior to the growth of light field research and 3D imaging in general, a lot of research has been done on evaluating the quality of 2D images. Quality evaluations can be divided into two classes - subjective and objective methods. Subjective methods are valuable because perceived quality is ultimately intended to reflect human perception and the human visual system is often more sensitive to certain aspects of quality than others. There are different procedures defined for performing experiments to evaluate subjective quality, such as Single Stimulus Continuous Quality Evaluation (SSCQE) and Double Stimulus Continuous Quality-Scale (DSCQS) specified by the International Telecommunication Union – Radiocommunication (ITU-R) standard for images [7]. However, obtaining meaningful results from subjective experiments is expensive and time-consuming because the test environment and procedure must be consistent. Objective methods, on the other hand, can be automated. Examples of well-known methods include the Video Quality Metric (VQM) [8], which takes into consideration additional aspects of the human visual system and statistical methods such as Peak-Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index (SSIM). Objective metrics can be categorized into Full Reference (FR), Reduced Reference (RR) and No Reference (NR) where FR metrics rely on complete information from a reference image and NR metrics derive quality from inherent attributes of the image [7].

There has been previous research to evaluate how the traditional 2D image quality metrics apply to 3D and light field content (which will be further described in Section 2). In terms of objective metrics, the typical approach has been to measure the PSNR or SSIM between each view of the reference light field image and the corresponding view in the processed image, followed by taking the average for the global metric [5]. However, light field content provides a lot of additional information, including structure across views, depth, and perspective. It is, therefore, worthwhile to create a quality metric that takes these inherent properties of light field content into consideration. In this paper, we propose an LF video quality metric that computes the horizontal, vertical, and diagonal EPIs of the reference and processed content to measure spatial and angular consistency. The major contribution of this paper is the inclusion of diagonal EPIs in the equation, which enables us to measure the angular consistency not only across horizontal and vertical views but also factors in the subtle angular changes introduced throughout the content.

The rest of the paper is organized as follows. Section II discusses the features and applicability of the state-of-the-art LF quality metrics. Section III presents the proposed method in detail. In Section IV, we discuss the experimental results. We conclude the paper in Section V.

II. RELATED WORKS

LF images are subject to a wide variety of distortions during acquisition, processing, compression, storage, transmission, and rendering; any of these steps may result in

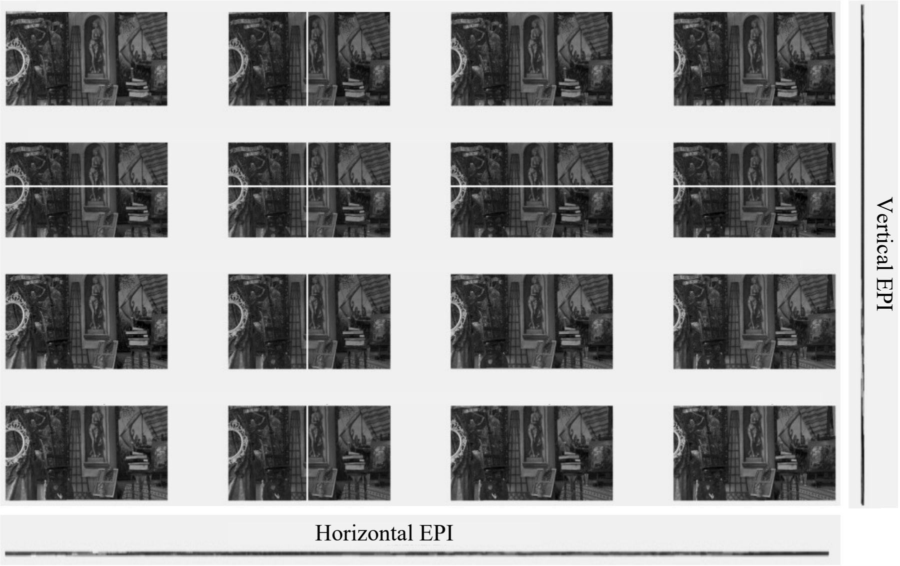


Figure 1. Horizontal and vertical EPIs of 'Painter' LF test sequence

visual quality degradation. The rapidly developing LF technology and consumer interest are pushing the need for objective quality evaluation of such contents.

One of the first works on LF quality evaluation by Adhikarla et. al [9] applies the traditional video metrics on individual light-field images and then averages the scores of overall images. They used Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure SSIM2D, which is widely used on 2D images, its extensions to angular domains – SSIM2D*1D, SSIM3D, and High Dynamic Range – Visible Difference Predictor (HDR-VDP-2), which stands out among perception-based quality metrics, the NTIA General Model for Video Quality Metric – VQM and the stereoscopic image quality metric – SIQM. To capture the full range of stereo quality metrics, they also included a stereoscopic video quality metric STSDLC. Another metric for the multiview video is MPPSNR, which computes the multi-resolution morphological pyramid decomposition on the reference and test images. Other metrics such as HDR-VDP-2, GMSD, STSDLC, and VQM perform well when comparing a distorted light field to a densely sampled reference LF. However, when a dense light field is not available, which is the case in camera array acquired LF, the usage of these metrics for quality assessment is not justified.

More recent works in LF quality metric domain, such as FR LFI-QA [10], measure the gradient magnitude similarity between the reference and processed content. Other methods [11] rely on depth maps. The accuracy of depth estimation, and consequently depth-map, depends on the method used and even with robust methods depth estimation is not always accurate. Hence, the proposed quality metric suffers inaccuracies too. [12] proposes an NR metric using horizontal and vertical EPIs to measure angular consistency. However, none of these methods fully exploit the structural and angular properties of LF. For example, for any given LF view [10] and [12] only consider the horizontal and vertical EPIs leaving the existing diagonal correlation underutilized. For dense LF content, this does not make a big difference as eventually the ray space is traversed twice. Though the relation is being indirectly factored into the quality metric using vertical and horizontal EPIs serially, for sparse LF content this indirect method cannot represent the quality accurately. Because of the wide baseline of the cameras and sparsely positioned cameras, we need additional scanning of the ray space to represent the quality accurately. Therefore, these methods are not suitable for sparse LF content.

III. Proposed Quality Metric

In order to design an LF quality metric for camera array-based (sparse) content, we leverage the horizontal, vertical, and diagonal EPIs. This way, we cover the ray space multiple times, and the quality metric can detect even the subtlest inconsistencies in the processed LF.

A. *LF and EPIs*

LF is described using a standard two-plane parameterization. Rays are defined using two parallel planes Π and Ω. The first plane Ω denotes image coordinates (x, y) ∊ Ω. The second plane Π contains the focal points (s, t) ∈ Π of all cameras. An entire 4D light field can thus be described by a function

$$R\ (s,t,x,y) \rightarrow L(s,t,u,v) \tag{1}$$

where $L(s,t,u,v)$ defines the intensity of the corresponding ray defined by the intersection (u,v) with the image plane and (s,t) with the focal plane, respectively. Furthermore, (u,v) can be treated as the spatial, and (s,t) can be treated as the angular resolution of the LF. Hence, there will be $s \times t$ sub-aperture views each having the resolution of $u \times v$. These sub-aperture views are slightly shifted from each other depending on the distance between the cameras. These disparities among the views can be estimated on the 2D slices

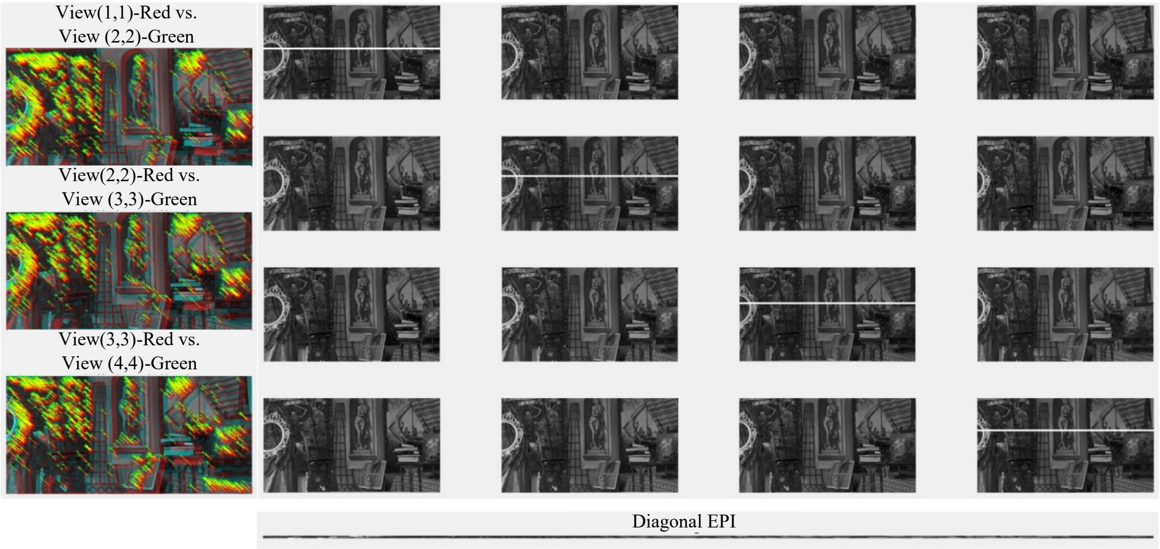


Figure 2. Diagonal EPI for views I_{1,1} ,I_{2,2} ,I_{3,3} , and I_{4,4} . The three figures on the left show an overlay of each pair of views along the diagonal and the matching Scale Invariant Feature Transform (SIFT) features that were detected.

$\sum_{t^*v^*} L$ from the 4D light field structure. This is achieved by setting t to a fixed value $t^*$and v to a fixed value v^* hence generating the EPIs.

$$S_{t^*v^*}\colon \sum_{t^*v^*} L \rightarrow R \qquad (2)$$

$$(u,s) \rightarrow S_{t^*v^*}(u,s) := L(s,t^*,v^*,u) \qquad (3)$$

Other slices with different fixed coordinates such as s and u, are defined analogously. Traditionally, EPIs have been used to estimate the depth and synthesize novel views. However, the EPIs can also be used to measure the spatial quality by comparing the original EPI and the processed EPI. The slopes created due to disparity among views can be measured to determine the angular quality of LF. To make the metric robust, we include the complete ray space multiple times in the form of horizontal, vertical, and diagonal EPIs.

B. *Generating Horizontal, Vertical & Diagonal EPIs*

The horizontal and vertical EPIs are generated as described in Shi et. al. [12]. Each view in the light field image is denoted as $I_{\{u,v\}}(s,t)$ where (s,t) is the spatial coordinate and (u,v) is the angular coordinate.

Each horizontal EPI is denoted as $E_{\{v^*,t^*\}}(s,u)$ where v^* and t^* are the fixed angular and spatial coordinates. Each row of the horizontal EPI contains the row of pixels from view $I_{\{u^*,v^*\}}(s,t^*)$. Similarly, the vertical EPI is denoted as $E_{\{u^*,s^*\}}(t,v)$. Each row of the vertical EPI contains the column of pixels from view $I_{\{u^*,v^*\}}(s^*,t)$.

Figure 1 shows an example of a horizontal and a vertical EPI. A total of 1088×4 horizontal EPIs and 2048×4 vertical EPIs are generated for each frame (all the views) of the LF.

The challenge with generating diagonal EPIs is to find the corresponding pixels in the diagonally aligned view. To determine which row of pixels in each view belongs to a particular diagonal EPI, the vertical offset is required. Therefore, we need to compensate for the shift experienced by the diagonal translation. For every pair of views along the diagonal, we detect all matching Scale Invariant Feature Transform (SIFT) features [13]. The matching features are used to calculate the average vertical offset and are used to determine the row of pixels in the next diagonal view to include in the EPI. To the best of our knowledge, this is the first work that uses SIFT to compensate for the shift in EPIs and create diagonal EPIs to measure angular consistency of LF content. Figure 2 shows an example of the matching SIFT features corresponding to each pair of views along the diagonal that contains $I_{\{1,1\}}, I_{\{2,2\}}, I_{\{3,3\}}$, and $I_{\{4,4\}}$. For the frame (all views) depicted in Figure 2, we have generated 28×2048 diagonal EPIs.

After generating all the EPIs, we have traversed the ray space three times and covered all the neighboring views of each view at least once. In this way, the quality metric can register even the subtle inconsistencies in the angular domain.

C. *EPI Similarity*

In order to quantify spatial quality, we compare the average PSNR and SSIM between the horizontal, vertical, and diagonal EPIs of the original and processed LF. This differs from the traditional quality metrics for LF content where the average PSNR or SSIM is calculated between each view separately and averaged to report quality. This method also provides insight into how the spatial relationship is maintained across the LF and applications requiring camera positions from content.

D. *EPI Gradient & Average Kurtosis*

We measure the angular distortion or deterioration using the pixel-wise gradient from horizontal, vertical, and diagonal EPIs. For each horizontal, vertical, and diagonal EPI, the gradient at each pixel is calculated using the Sobel

TABLE I. LF SPATIAL QUALITY FROM EPI SIMILARITY

PSNR					
	QP45	**QP40**	**QP35**	**QP30**	**QP25**
Horizontal EPI	30.64	33.14	35.64	37.89	39.95
Vertical EPI	30.91	33.43	35.90	38.07	40.04
Diagonal EPI	29.21	31.37	32.95	33.80	35.39
Average PSNR	**30.26**	**32.65**	**34.83**	**36.59**	**38.46**
SSIM					
	QP45	**QP40**	**QP35**	**QP30**	**QP25**
Horizontal EPI	0.89	0.94	0.96	0.97	0.98
Vertical EPI	0.88	0.93	0.95	0.97	0.98
Diagonal EPI	0.85	0.90	0.92	0.94	0.96
Average SSIM	**0.88**	**0.92**	**0.95**	**0.96**	**0.97**

gradient operator. A histogram of the gradients is generated, and we use the average kurtosis as a metric to describe the amount of distortion introduced from compression.

IV. EXPERIMENTS AND EVALUATION

LF content can experience degradation from different operations, such as compression, calibration, super-resolution, and other operations. In this paper, we validate the proposed quality metrics by using LF videos compressed at different Quantization Parameter (QP) levels of 25, 30, 35, 40, and 45 using the MV-HEVC, Multi-view extension of High Efficiency Video Coding [14], to compress light field content. We use camera array-based LF test sequences [6]. We report the results for the 'Painter' test sequence in this paper. Other test sequences show consistent performance.

TABLE II. KURTOSIS OF GRADIENT DIRECTION HISTOGRAM

	Reference	**QP45**	**QP40**	**QP35**	**QP30**	**QP25**
Horizontal EPI	1.36	1.23	1.26	1.27	1.29	1.30
Vertical EPI	1.40	1.24	1.26	1.28	1.29	1.31
Diagonal EPI	1.37	1.24	1.26	1.28	1.29	1.31
Average Kurtosis	**1.38**	**1.24**	**1.26**	**1.28**	**1.29**	**1.30**

Table 1 contains the average PSNR and SSIM values between all horizontal, vertical, and diagonal EPIs from the compressed LF with respect to the original video. The results indicate that the EPI similarity correlates with different amounts of compression.

The pixel-wise gradient calculation is presented in Figure 3 in 10° bins (36 bins in total). We can adjust the number of bins depending on the level of accuracy desired. We can quantify the results of gradient direction using kurtosis. Table 2 contains the average kurtosis values for all horizontal, vertical, and diagonal EPIs. The results show that with increasing QP levels or more compression, the kurtosis of the gradient histogram decreases. Figure 3 illustrates this relationship. The gradient histograms show the count of

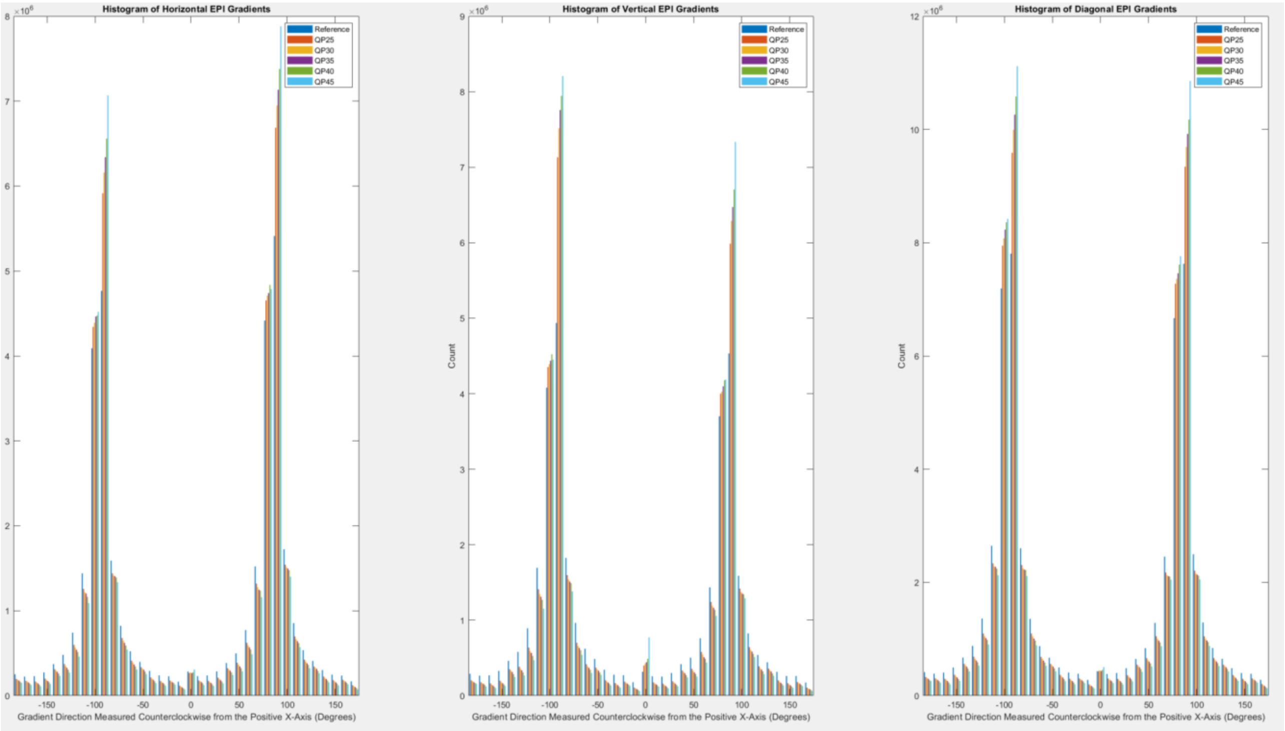


Figure 3. Gradient histograms of the horizontal, vertical, and diagonal EPIs. The histograms show that the gradient direction is more concentrated at ±90° for higher QP values.

gradients from all EPIs and are presented in bins of $10°$. The angle is measured counterclockwise from the positive $x-axis$. For each $10°$ bin, the histograms show the gradient counts at each QP level. The histograms show that with increasing QP levels, the gradients become more concentrated at $\pm 90°$, which is likely a result of the stepwise artifacts introduced in the EPI with more compression. Since the histogram is centered at $0°$, it makes sense for the kurtosis to decrease with more compression since the tails of the histogram at $\pm 90°$ are larger relative to the center.

V. Conclusion and future work

In this paper, we proposed an LF quality metric using horizontal, vertical, and diagonal EPIs. Our major contribution was the SIFT-assisted diagonal EPI translation for scanning the diagonal ray space of the LF. Such a quality metric is useful in determining the spatial and angular consistency of processed LFs with the original. We applied our metrics to compressed LFs and found the metric can accurately describe the quality of a sparse LF. An important next step would be further validation of our metrics by evaluating LFs that have undergone other distortions such as super-resolution, calibration, etc. We also intend to correlate our results with subjective testing scores.

References

[1] M. Levoy and P. Hanrahan, “Light Field Rendering,” In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996, Aug 1, pp. 31-42.

[2] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light Field Photography with a Hand-held Plenoptic Camera,” 2005 (Doctoral dissertation, Stanford University) .

[3] C. Zhang, G. Hou, Z. Zhang, Z. Sun, and T. Tan, “Efficient auto-refocusing for light field camera,” Pattern Recognit., vol. 81, pp. 176–189, Sep. 2018, doi: 10.1016/j.patcog.2018.03.020.

[4] C. Conti, L. D. Soares, and P. Nunes, “Light field coding with field-of-view scalability and exemplar-based interlayer prediction,” IEEE Trans. Multimed., vol. 20, no. 11, pp. 2905–2920, 2018, doi: 10.1109/TMM.2018.2825882.

[5] N. Mehajabin, M. T. Pourazad, and P. Nasiopoulos, “An Efficient Pseudo-Sequence-Based Light Field Video Coding Utilizing View Similarities for Prediction Structure,” IEEE Trans. Circuits Syst. Video Technol., 2021, pp. 1-16, doi: 10.1109/TCSVT.2021.3092282.

[6] N. Sabater et al., “Dataset and Pipeline for Multi-view Light-Field Video,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2017-July, pp. 1743–1753, 2017, doi: 10.1109/CVPRW.2017.221.

[7] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, “Quality analysis for 3D video using 2D video quality models,” IEEE Trans. Consum. Electron., vol. 54, no. 4, pp. 1969–1976, 2008, doi: 10.1109/TCE.2008.4711260.

[8] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando, and A. M. Kondoz, “Perceptual video quality metric for 3D video quality assessment,” 3DTV-CON 2010 True Vis. - Capture, Transm. Disp. 3D Video, pp. 1–4, 2010, doi: 10.1109/3DTV.2010.5506331.

[9] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, and P. Didyk, “Towards a quality metric for dense light fields,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2, 2017, pp. 58–67.

[10] Y. Fang, K. Wei, J. Hou, W. Wen, and N. Imamoglu, “Light Filed Image Quality Assessment by Local and Global Features of Epipolar Plane Image,” 2018 IEEE 4th Int. Conf. Multimed. Big Data, BigMM 2018, pp. 1-6, doi: 10.1109/BigMM.2018.8499086.

[11] P. Paudyal, F. Battisti, S. Member, M. Carli, and S. Member, “Reduced Reference Quality Assessment of Light Field Images,” IEEE Trans. Broadcast., vol. 65, no. 1, pp. 152–165, 2019.

[12] L. Shi, W. Zhou, Z. Chen, and J. Zhang, “No-Reference Light Field Image Quality Assessment Based on Spatial-Angular Measurement,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 11, pp. 4114–4128, 2020, doi: 10.1109/TCSVT.2019.2955011.

[13] D. G. Lowe, “Object recognition from local scale-invariant features,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2, pp. 1150–1157, 1999, doi: 10.1109/iccv.1999.790410.

[14] G. Tech, Y. Chen, K. Müller, J. R. Ohm, A. Vetro, and Y. K. Wang, “Overview of the multiview and 3D extensions of high efficiency video coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 1, pp. 35–49, 2016, doi: 10.1109/TCSVT.2015.2477935.

Current Status of Examples of Initiatives Using Open Data in Government

Participating in the First Governor's Cup Open Data Hackathon in Tokyo

Yusuke Takamori, Junya Sato, Masahiro Fujimoto
Electronic Information Course
Polytechnic University
Kodaira-shi, Tokyo
e-mail: b19309@uitec.ac.jp, b19308@uitec.ac.jp, b19315@uitec.ac.jp

Masaki Endo, Shigeyoshi Ohno
Division of Core Manufacturing
Polytechnic University
Kodaira-shi, Tokyo
e-mail: endou@uitec.ac.jp, ohno@uitec.ac.jp

Daiju Kato
Nihon Knowledge Co. Ltd.
Taito-ku, Tokyo
e-mail: d-kato@know-net.co.jp

Hiroshi Ishikawa
Graduate School of Systems Design
Tokyo Metropolitan University
Hino-shi, Tokyo
e-mail: ishikawa-hiroshi@tmu.ac.jp

***Abstract*—This paper reports on services using open data that we proposed at the "First Governor's Cup Open Data Hackathon" held by the Tokyo Metropolitan Government. The use of open data is necessary for the digitization of public administration. We participated in a hackathon organized by the Tokyo Metropolitan Government, developed an application that allows users to search for and participate in event information using open data, and proposed it as a solution for regional revitalization. This proposal demonstrated the usefulness of open data and its potential for solving administrative issues. Nevertheless, open data in Tokyo remains underdeveloped: it needs improvement. We will also report on improvements we made through this hackathon.**

Keywords–big data; civic tech; Hackathon; open data.

I. INTRODUCTION

Open data are held by the national government, local governments, private companies, etc. The data are made publicly available on the Internet so that anyone can use the data freely and free of charge. Secondary use of open data by the public and by companies is permitted. In addition, the data must be in a format that is easily read by computers. In Japan, the Basic Act on the Promotion of Public-Private Data Utilization [1] was enacted in 2016 to promote the use of public–private data.

Approximately 96% of prefectures and municipalities responded to the "Questionnaire on Open Data Initiatives for Local Governments" [2] administered by the Information and Communications Technology (IT) Strategy Office, Cabinet Secretariat in 2020. Of the organizations which responded (prefectures (47 responses) and municipalities (1,668 responses)), 56.6% had already released open data. The top three open data items released to the public were basic statistical information (population, industry, etc.) at 37.7%, various information in the disaster prevention field (designated emergency evacuation sites) at 41.0%, and information related to the location of public facilities and services (list of public facilities and list of administrative services) at 33.6%. However, the top three challenges and problems in working with open data were unclear effects, benefits, and needs of open data (50.6%), lack of human resources in charge of open data (55.4%), and lack of progress in the utilization of open data (29.2%). As indicated by survey results, open data are becoming increasingly available to the public in government. However, the lack of human resources in charge of open data and the lack of understanding of how to use open data have not promoted initiatives related to open data in Japan.

In this context, the Tokyo Metropolitan Government, led by its Digital Services Bureau, strengthened its efforts to use open data. As of December 2021, the Tokyo Metropolitan Government has released more than 49,000 items of open data [3], about 70% of which are in computer-readable Comma Separated Values (CSV) or other formats. The Tokyo Metropolitan Government, the most open data-oriented government in Japan, still has too few initiatives for open data utilization. For this reason, the Tokyo Metropolitan Government held the first Governor's Cup Open Data Hackathon as a five-day event in December 2021. This event was an effort to solve Tokyo's administrative issues and to use civic technology to implement new social digital services that are expected to improve Tokyo residents' quality of life. The Hackathon received 68 applications for participation from 186 people.

This paper reports on issues related to open data in Tokyo, based on services we proposed and prototyped at the first Governor's Cup Open Data Hackathon [4] and based on findings from the hackathon.

The remainder of the paper is organized as follows. Section 2 presents earlier research related to this topic. In Section 3, we present our proposed use of open data for public administration. Section 4 describes the open data used in the proposed methodology. Section 5 describes the prototype created using the proposed method. Section 6 describes the challenges found in the process of creating the prototype. Section 7 describes the future of open data in government.

II. Related Work

Various applications using open data have been developed. Although the number of applications in Tokyo is small, there are some published examples [5]. The Fuchu Barrier-Free Restroom MAP [6] visualizes barrier-free restrooms in Fuchu City using the Tokyo Metropolitan Open Data Catalog. In addition, the River Camera Dashboard [7] application, a one-stop dashboard, visualizes the locations and transmitted images of river-monitoring cameras managed by the Bureau of Construction of the Tokyo Metropolitan Government and sea surface live cameras managed by the Bureau of Ports and Harbors of the Tokyo Metropolitan Government. "LinkedSpending: OpenSpending becomes Linked Open Data" [8] presents a platform for using government spending as open data in countries around the world. Also, "Interactive Navigation of Open Data Linkages" [9] shows a web application for utilizing open data.

Next is an example of a hackathon conducted by governments. Few examples exist in Japan, but hackathons use civic technology in other countries. The President's Cup Hackathon [10] in Taiwan, to which the Tokyo hackathon referred, has been held since 2018. The international track in that hackathon is open to international participants. Hack the Crisis is a hackathon proposed by Accelerate Estonia, an Innovation Lab led by the Estonian government, and organized by Garage48 [11], a hackathon management organization, to fight the new coronavirus. Approximately 1,300 people from more than 20 countries participated in this online hackathon. An operations manual was also published. The event became a worldwide event, especially in Europe.

Consequently, hackathons that use open data and civic technology are becoming popular. In Japan, a need exists for activities using digital technology to promote available data by bringing together those who use data, such as companies and citizens, and those who own the data, such as governments.

III. Proposed Service

We developed an application to help revitalize local communities in Tokyo using open data. In Tokyo, the declining birthrate, aging population, and influx of people from rural areas have diluted local communities and have become an administrative issue. Furthermore, in recent years, the Corona pandemic has reduced the number of neighborhoods and has accelerated the weakening of local communities. In addition, community ties are an essential element of smooth communication during a disaster and are necessary for an earthquake or fire. We responded to this challenge by creating an application that provides an opportunity to develop local communities using open data.

This proposal will be developed as a smartphone application that anyone can use. The smartphone application allows users to find and participate in data on events that are the first step in building community connections. The application is divided by ward, city, town, and village, allowing users to browse events in the neighborhoods where they live. The application targets elderly people, people who are moving to Tokyo for higher education or employment, and singles in households. By helping them find new places to live, we can create a comfortable living environment for Tokyo residents.

IV. Used Open Data

Widely diverse open data are available, including data of local stores and public facilities. Among them, we chose event data as the target for utilization. Event data include information related to events held by local communities: the data are more relevant to issues of regional revitalization than other related data. The Tokyo Metropolitan Government publishes open data on events in each ward, city, town, and village. Event information is uploaded to the catalog site by each ward, city, town, and village; the information is released as open data. We confirmed that 17 wards, cities, towns, and villages published event information related to the Tokyo Metropolitan Government's portal site.

Information related to events in various municipalities is available, but we used open data of Koto-ku as a demonstration, mainly for two reasons. The first is the abundance of information. Unlike other wards, cities, towns, and villages, Koto-ku has large amounts of information related to events. We were able to confirm various information about the possibilities. We decided to use Koto-ku data because we judged that this amount of detailed event information was sufficient to realize our application. The second reason was the data precision. Most open data published by the Tokyo Metropolitan Government included large amounts of null information. Among them, Koto-ku was selected as a demonstration because it has few nulls and

TABLE I. Open data Content of the Event

Category	Description
Place	Prefectures, Municipality, Address, Latitude, Longitude, How to access, Postcode, Distance
Event	Event name, Event details, Description
Date and time	Start date and time, End date and time, Start time, Ending time
Contact	Contact name, Phone number
Organizer	Organizer name
Participation information	How to apply for participation, Description of the deadline date and time information, Participation Conditions
Home page	Description of homepage Uniform Resource Locator (URL), Event image
Capacity	Description of capacity by age
Remarks	Presence or absence of parking lot, Presence or absence of a nursery center and details

large amounts of event information, which we judged to be easy to handle for our use. For these reasons, we created an application using event information of Koto-ku as the first demonstration data among the data of many wards, cities, towns, and villages.

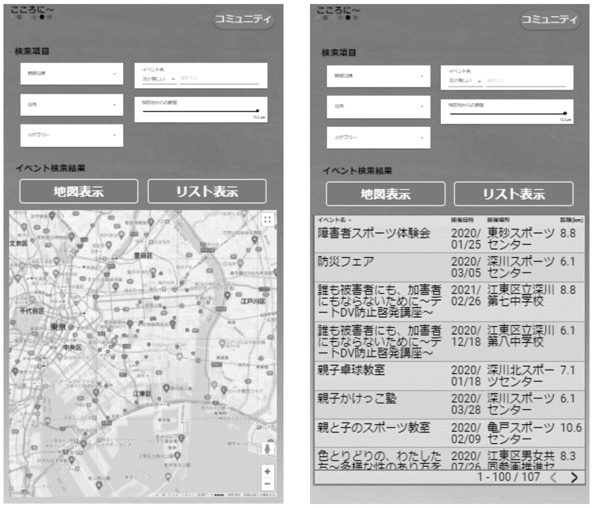
We obtained data for Koto-ku from the Tokyo Metropolitan Government's Open Data Portal site. The open data for Koto-ku events included 107 event data, and included 55 pieces of detailed event information. Table I shows typical events, put into nine categories: location, event, date and time, contact information, organizer, participation information, website, capacity, and remarks. The nulls, representing missing data among these open data, were 911 out of 5564 pieces of total data, or approximately 17% of the total. Koto-ku event data had the lowest level of missing data among the 17 wards and municipalities in the Tokyo portal.

V. PROTOTYPE

To provide services using open data, we created a prototype of an application. The prototype implemented the functions and User Interface (UI) to search for and participate in Tokyo Metropolitan Government event information. We developed the prototype as a smartphone application. Smartphones are currently gaining popularity rapidly in Japan. As of May 2021, the smartphone penetration rate in Japan is 92.8%. To enable more people to use them, we chose a smartphone application because it is accessible to everyone.

The prototype of this application targets Koto-ku, Tokyo. Its essential functions are to search for event information in the area to obtain details and to participate in events. It is also necessary to maintain the new community created through the event. For this reason, we implemented a chat function within the community. The application consists of a screen for seeking event information to find a suitable event, a screen displaying detailed information about each event, and a screen for chatting with other event participants.

Figure 1 shows the UI of the screen intended for searching for event information and finding an event. We divided this screen into search items and event search results. The event information matches the conditions entered in the search items displayed immediately in the results. In this prototype, to search for necessary information from the vast amounts of data, we added a function that allows users to search by any of the following items: date, time, address, category, keyword, and distance. By narrowing down event information from multiple criteria, one can obtain the exact information users want from a vast amount of data. Additionally, we implemented a function to switch between list (Figure 1a) and mapping (Figure 1b) formats to display results of retrieved event information. By making it possible to switch the display format, the UI design facilitates the comparison of items of importance according to user needs.



イベント名	開催日時	開催場所	距離(km)
障害者スポーツ体験会	2020/ 01/25	東砂スポーツ センター	8.8
防災フェア	2020/ 03/05	深川スポーツ センター	6.1
誰も被害者にも、加害者にもならないために～デートDV防止啓発講座～	2021/ 02/26	江東区立深川 第七中学校	8.8
誰も被害者にも、加害者にもならないために～デートDV防止啓発講座～	2020/ 12/18	江東区立深川 第八中学校	6.1
親子卓球教室	2020/ 01/18	深川北スポー ツセンター	7.1
親子かけっこ塾	2020/ 03/28	深川スポーツ センター	6.1
親と子のスポーツ教室	2020/ 02/09	亀戸スポーツ センター	10.6
色とりどりの、わたしたち～多様な性のあり方を	2020/ 07/26	江東区男女共 同参画推進セ	8.3

(a) Display in map format (b) Display in list format

Figure 1. Event information search screen.

The screen displaying detailed information for each event is displayed when the event in Figure 1 shown earlier is selected. It consists of detailed information for that event. The detailed information includes the location, date and time, event details, contact information, and other information required by event participants. Event participants can chat freely with other participants. We generated a chat room for each event. The chat room can be accessed from event details. Participants are provided a place for participants in the same event to communicate.

This application realized open data by providing the information desired by users in a visual, easily understandable form. We also believe that using open data will help revitalize the community.

VI. RESULTS OF WORKING ON THE HACKATHON

We used open data of event information published in only 17 of the 62 wards, towns, and villages in Tokyo. The percentage of wards, towns, and villages that publish open data was 27%. Of the 17 wards, cities, towns, and villages which publish open data on events, 15 publish open data in CSV files. The remaining two publish open data in different formats. They updated only one within one month of the 15 wards, towns, and villages which published their open data in CSV files. Consequently, the open data released by the Tokyo Metropolitan Government are of various formats. The stated update frequency is not reliable, making it difficult for users to use the data. For future data use, we hope that the Tokyo Metropolitan Government will create a unified standard for data to be released using the recommended dataset of the Government Chief Information Officers (CIO) Portal and other data and create mechanisms to increase the update frequency.

The prototype we created was intended to solve the problem by providing open data information to users in an easily searchable format. However, in terms of using open data, it is preferable to process event information into an easy-to-read format and to provide it in combination with open data of public facilities and generate new data and value by making predictions and inferences from the original data before providing the data. We regard the true meaning of using open data as providing newly created value that has not been available until now, only from event information by combining event information and information of other types.

Administrative issues include employment, public safety, air pollution, and many others. To resolve administrative difficulties, we use open data to leverage the digital skills of residents. This goal necessitates enhancement of open data that the government makes available free of charge in a form that is useful for secondary purposes. This event confirmed that open data help resolve administrative issues.

VII. Conclusion

This paper reports services using open data proposed by the Tokyo Metropolitan Government in the first Governor's Cup Open Data Hackathon held by the Tokyo Metropolitan Government. For this effort, developing an application that will become a service through the utilization of open data and civic technology is also underway. Consequently, the hackathon conducted by the government demonstrated the possibility of solving administrative issues. However, results also clarified that a need exists for measures to enhance open data in public administration and to create a mechanism to involve more engineers with digital skills in the digitization of public administration. Our future reports will present proposals to the administration, including points for improving the Tokyo Metropolitan Government's open data, which we have learned through our research using open data.

Acknowledgment

This work was supported by JSPS KAKENHI Grant No. 18K13254. This work was supported by JSPS KAKENHI Grant Number 20K12081, Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas, and the Okawa Foundation Research Grant.

References

[1] Prime Minister of Japan and His Cabinet. *Basic Law for the Promotion of Public-Private Data Utilization*. [Online]. Available from: https://www.kantei.go.jp/jp/singi/it2/hei/detakatsuyo_honbho.html 2016.12.07 [retrieved: April, 2022]

[2] Government CIOs' Portal, Japan. *Results of Questionnaire on Open Data Initiatives for Local Governments*. [Online]. Available from: https://cio.go.jp/sites/default/files/uploads/documents/r2_survey_results.pdf 2021.06.09 [retrieved:April, 2022]

[3] Tokyo Metropolitan Government. *Tokyo Metropolitan Government Open Data Catalog Site*. [Online]. Available from: https://catalog.data.metro.tokyo.lg.jp/dataset 2017.03.24 [retrieved:April, 2022]

[4] Tokyo Metropolitan Government. *Tokyo Governor's Cup Open Data Hackathon*. [Online]. Available from: https://portal.data.metro.tokyo.lg.jp/hackathon/ 2021.11.05 [retrieved:April, 2022]

[5] Tokyo Metropolitan Government. *Tokyo Governor's Cup Open Data Application examples, etc*. [Online]. Available from: https://portal.data.metro.tokyo.lg.jp/case/ 2021.6.28 [retrieved:April, 2022]

[6] Tokyo Metropolitan Government. *Fuchu Barrier-Free Restroom MAP*. [Online]. Available from: https://portal.data.metro.tokyo.lg.jp/case/case-study-01/ 2021.6.28 [retrieved:April, 2022]

[7] Tokyo Metropolitan Government. *@KentoIDE: River Camera Dashboard*. [Online]. Available from: https://portal.data.metro.tokyo.lg.jp/case/case-study-07/ 2021.11.11 [retrieved:April, 2022]

[8] H. Konrad, M. Michael and L. Jens, “LinkedSpending: OpenSpending becomes Linked Open Data,” Semantic Web, vol. 7, pp. 95-104, Jan. 2016, doi: 10.3233/SW-150172.

[9] Erkang Zhu, Ken Q. Pu, Fatemeh Nargesian and Renee J. Miller, “Interactive Navigation of Open Data Linkages,” Proceedings of the VLDB Endowment, vol. 10, pp. 1837–1840, Aug. 2017, doi: 10.14778/3137765.3137788.

[10] Taiwan. *2021 Presidential Hackathon*. [Online]. Available from: https://presidential-hackathon.taiwan.gov.tw/en/ 2021.6.24 [retrieved:April, 2022]

[11] Garage48. *Hack the Crisis*. [Online]. Available from: https://garage48.org/hackthecrisis 2020.3.13 [retrieved:April, 2022]