Presenter-Centric Image Collection and Annotation: Enhancing Accessibility for the Visually Impaired

Luísa Ferreira*[†], Daniel Fernandes[†], Fabio Cerqueira[‡], Marcos Ribeiro[†], and Michel Silva[†],
[†]Department of Informatics, Universidade Federal de Viçosa - UFV, Viçosa, Brazil.
[‡]Department of Production Engineering, Universidade Federal Fluminense - UFF, Petrópolis, Brazil
{luisa.ferreira*, daniel.louzada, marcosh.ribeiro, michel.m.silva}@ufv.br, frcerqueira@id.uff.br

***Abstract*—With the recent COVID-19 pandemic, our daily lives and workplace became more dependent on technologies, *e.g.*, the intensive use of video conferences. While a highly connected world enables remote working, it also raises new barriers for an already excluded community, the visually impaired people. Low or no eyesight prevents those people from capturing visual information, which makes it difficult to understand the overall context of a remote presentation. Most Artificial Intelligence methods are specific to sighted-people data domains due to the scarcity of datasets for the visually impaired. In this paper, we propose an approach to collect data automatically and a protocol to annotate this data specifically for this audience, aiming to support the development of Assistive Technology systems. We demonstrate the viability of the proposed methods by creating a dataset and evaluating its quality, diversity, and representativeness through analytical methods and machine learning models.**

I. Introduction

Guiding the advancements in Artificial Intelligence (AI), a significant number of datasets from diverse domains have been generated. The combination of datasets and machine learning models revolutionized many fields by providing a successful accomplishment of real-world tasks and releasing tools present in our daily lives. Noteworthy examples include improvements in the precision agriculture field [1], real-time object detection [2], [3], and estimation of COVID-19 contamination risk based on social distancing and face mask detection [4].

The AI revolution also targets Assistive Technology, *e.g.*, detection of dangerous object [5], Image Captioning [6], [7], and Visual Question Answering systems for visually impaired people [8]. Despite the impressive results, those methods are bound by specific tasks and data domains. Individuals with visual impairments still face countless daily challenges, ranging from recognizing the object they interact with to understanding the context of multimedia news [6], [9].

The COVID-19 pandemic was marked by significant transformations, and boosted our dependence on technology [4]. Some of those transformations became widespread nowadays, *e.g.*, home office as a substitute for commuting and desk work, and webinars as a successor for in-person conferences, interviews, meetings, or even classes. In this highly connected world, the visually impaired struggle to socialize and understand the context of shared content. Visual information is complementary to the message itself, *e.g.*, a social distance warning from a front-line health member spoken from a crowded hospital is more sensitive than the same warning from business people in their skyscraper office. People suffering from low vision are deprived of this visual context and their understanding relies entirely on the non-visual information.

With the technological boom triggered by the pandemic, a trend towards inclusion has also emerged on social media, *e.g.*, the Brazilian project #ForBlindToSee, which aims to provide image description by requesting video conference presenters to describe themselves before the talk so as to include the low vision audience. The description usually covers the main characteristics of the presenter's appearance and surroundings. The drawback of this initiative is the fact that it is made manually and by sighted people, which means that the description is frequently missing or poorly informative.

Although some efforts have been proposed to automate this description process and remove the need for sighted people to describe the image [6], [8], [9], the main barrier to this task is the data domain, since images captured by blind people do not follow the same distribution of massive general purpose datasets [5], [7], [8]. Furthermore, the image descriptions for this audience do not follow the same rules applied to the dataset created for sighted people [10], [11]. Therefore, the need for datasets targeting supervised models is imperative.

This work presents an approach to gather images from shared content featuring a single foreground presenter. The purpose of this approach is to construct a dataset of webinar images aiming to support computer vision tasks related to webinars specifically for low eyesight people. In addition to the image gathering strategy, we also developed an annotation protocol for the task of manually describing image visual context for this specific audience. We make public a dataset[1] with 10,939 images, from which 967 are annotated, following the proposed gathering and annotation protocols.

II. Related Work

In this section, we highlight the relevance of tailored datasets for real-world edge AI applications.

Besides the general-purpose and massive datasets proposed by big corporations or consortia of top-tier research institutes, specific datasets are still necessary to address local or regional real-world applications. Kuhn *et al.* proposed the BRCars, a

[1] Link: https://github.com/MaVILab-UFV/presenter-centric-dataset-SIBGRAPI-2023

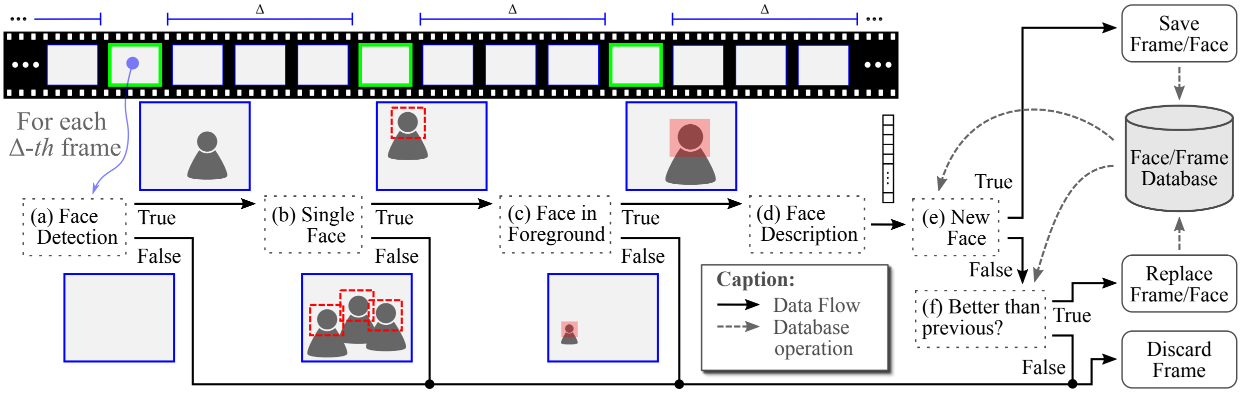

Fig. 1: Overview of the data gathering process for each input video. For each frame, we analyze the presence of a face (a), if the face is unique (b) and of a foreground presenter (c). If any check returns false, the frame is dropped. Otherwise, we describe the face (d) to check if it is the first appearance (e). In affirmative case, we save the face. In negative case, we replace the cached version with the current face if the current is better than the cached version.

dataset comprised of $\sim 2{,}000{,}000$ images to aid the classification task on the Brazilian car domain [12]. Nascimento *et al.* propose a dataset with a thousand images of soybean seedlings with annotation to the genotype and soil condition to aid the development of agriculture AI models [1].

As we witness exponential growth in both number and quality of general-purpose AI applications, much slower progress is observed for specific tasks. Also, there are few publicly-available datasets targeting problems of minorities, such as visually impaired people [2], [8].

In the context of object recognition, the Teachable Object Recognizer (TOR) task tends to improve visually impaired lives by learning new object classes from only a few samples. The ORBIT dataset with $4{,}733$ videos of 588 objects was proposed to enable TOR in real-world scenarios specifically for visually impaired people [3]. Recently, Tang *et al.* introduced a dataset with $7{,}915$ images from 15 types of common obstacles in the path of low-vision people, targeting models focused on enabling traversability in outdoor environments [2].

Image Captioning techniques have the potential to serve as Assistive Technologies by generating textual descriptions of images [6], [7]. However, they produce generic and vague descriptions. Dense Captioning [13] and Image Paragraph generate more detailed descriptions, but those are not suitable for the visually impaired, since these methods tend to produce unorganized long sentences or fail to focus on the relevant characteristics of interest [9]. Gurari *et al.* proposed a dataset for Image Captioning tasks focusing on the interests of visually impaired people [6]. The dataset consists of $39{,}000$ photos captured by blind people and textually described with 5 captions. The same group proposed a dataset with $5{,}537$ photos to boost the development of algorithms to address the problem of unintended disclosures of private information [7].

The Visual Question Answering (VQA) task has great potential to assist blind individuals in their daily activities. However, many VQA methods encounter difficulties in handling low-quality images. Aiming to encourage the rise of solutions for this problem, the VizWiz-VQA dataset was proposed with images captured by blind people in real-world scenarios with a related question and 10 annotated answers for each image [8].

The last cited datasets are composed of images captured by the visually impaired. In addition, Shah *et al.* created a dataset of $2{,}500$ manually photographed and web-scrapped images to train a model to identify 5 classes of dangerous objects, aiming to develop an alert system for people with visual disabilities [5]. Similarly to Shah *et al.*, in this study, we proposed a dataset to encourage the development of methods to aid blind people to perceive the world. We focus on the problem of extracting context from webinar images specifically for people with low vision.

III. Dataset

In this section, we present the data collection process and the annotation protocol to create the dataset.

A. *Data Gathering*

For each input video, we analyze every Δ-*th* frame and create an empty database to store frames metadata, which includes frame id, face descriptor, and face detection confidence. This Δ factor is used to avoid processing frames with very similar content that leads to a high probability of being the same presenter. The first step is to apply a face detection model to determine if there is a face on the frame (Fig. 1-a), followed by a verification if only a single individual was identified (Fig. 1-b), since our method targets on images with only one presenter.

Next, we check if the detected face is in the foreground (Fig. 1-c) by comparing if its area is larger than $\gamma\%$ of the frame area. An affirmative return indicates that the face is of a noteworthy person to describe. If any of the three checks returns false, we drop the frame.

After a successful frame extraction, we describe the identified face (Fig. 1-d) and use the created descriptor to verify if this is its first occurrence (Fig. 1-e). If this is the case, the frame metadata is written to the frame database. This prevents saving multiple frames of the same person within the same video, since more than one annotation of the same presenter in the same video would be redundant. More specifically, this step checks if the cosine distance between the face descriptor and each descriptor stored in the frame database is less than a threshold α. If one or more faces in the database achieve a distance less than the threshold α, it means that the current person could be already detected. To ensure accuracy, we also employed the Euclidean distance between the potential previous appearance and the current frame, and if the value is less than a threshold β, it indicates that the current person was already identified in the video and stored in the database.

In cases where the face was already detected in the video, we compare the face detection confidences of both the frame being analyzed and the previously recorded frame, and keep only the one with the highest confidence (Fig. 1-f).

B. Annotation Protocol

We proposed a protocol for describing the visual context of images, specifically for the visually impaired people, based on the guidelines and good practices recommended by the Brazilian Association of Technical Standards [10] and the Perkins School for the Blind [11]. Those entities provide guidelines for describing audio-visual data aiming to promote awareness and understanding for people with low vision.

The guidelines suggest the creation of descriptions that include characteristics of interest of both physical and visual appearance of individuals, *i.e.*, gender, age group, ethnicity, skin, hair and eye color, facial expressions, and other notable features. Additionally, descriptions should mention clothing, accessories, objects, and scene [10]. Even though blind people may have never seen colors, they often have knowledge of color associations and it can thus aid their understanding [11]. A good practice to build a description is to assure that the resulting text is precise, simple, coherent, fluid, in the present tense and within the range of 125 to 280 characters [10], [11].

Once rules were set, we performed a two-step annotation task conducted by three human annotators who were presented with the image description rules for the visually impaired. In the first step, annotators utilized custom annotation software to describe each image in a paragraph format, adhering to the guidelines and best practices outlined in the set of rules. In the second step, a description cross-adjustment was performed, in which an annotator reviewed an image along with its description previously annotated by one of the other two annotators. The current annotator was then tasked with adjusting the description according to the protocol. The cross-adjustment process was conducted blindly and randomly, involving potential additions, edits, or removals to the description. As a result, each annotated image yielded three descriptions: the original description and two adjusted descriptions. The descriptions were initially written in Brazilian Portuguese and subsequently translated into English using a translation tool.

C. Dataset Construction

With the aim of creating a dataset composed by diverse and representative images containing a single foreground presenter, we selected a set of publicly available YouTube playlists with such content, *e.g.*, webinars, news, interviews, online classes, tutorials, *react* videos, podcasts, and similar ones. For each playlist video, we applied the process described in Sec.-III-A, resulting in a frame database and a set of frames. After processing all videos from the selected playlists, we organized the frames in three datasets, as detailed next.

a) Complete Dataset: This version of the dataset represents the union of selected frames from all videos of all playlists. It is worth mentioning that for each video, only one image is selected for each person. However, the presence of a significant number of videos featuring programs with a fixed host (*e.g.*, news) leads to multiple images of the same person. Although those images came from different videos, there are only slight variations in clothing and settings. The Complete Dataset contains 10,939 images taken from 4,173 videos.

b) Single Person Dataset: Aiming to increase the data diversity, we proposed a dataset version in which each person appears only once. To achieve this, we create a global frame database by concatenating the frame database of the current processed video with the frame databases of previously processed videos. Then, we only insert the selected frame if the person is not found on the global database. For this check, we employed the process described in Sec. III-A. The proposed version, by not inserting frames from people who are already in the global frame dataset, contains 5,689 images when processing the same 4,173 videos of the Complete Dataset.

c) Annotated Single Person Dataset: To generate this version of dataset, we randomly select images from the Single Person Dataset and apply the process described in Sec. III-B to label the images. The selected images are equally distributed to the annotators for the annotation task. For the cross-adjustment task, each annotator adjust all descriptions from the other annotators. This process took 5 months, and the publicly available version of the dataset contains 684 images with three descriptions, 190 images with two descriptions, and 93 images with a single descriptions, totaling 967 annotated images. Fig. 2 depicts dataset samples exemplifying the representativeness and diversity of the data.

IV. Data Analysis

We evaluate the quality, diversity, and representativeness of the created dataset using multiple analytical methods and machine learning models aiming to mitigate bias.

A. Image Quality

We analyze the low-level features of the images, *e.g.*, brightness, contrast, and sharpness, and high-level features, such as the number of objects. To measure the image sharpness, we applied the Tenengrad method that calculates the variance of the resulting gradient of the grayscale version of the image. Regarding contrast analysis, we calculate the ratio of the brightest pixel gray level to the darkest. The brightness

Description: A Caucasian woman with a cheerful facial expression. She is wearing red blouse, lipstick and earphone. The woman has short brown straight hair. At the bottom there is a shelf with objects, a chandelier and a vase of flowers.

Adjust #1: A Caucasian woman with a cheerful facial expression. She is wearing red lipstick, earrings, headphones and she is wearing blouse with red collar. The woman has short brown straight hair. At the bottom there is a shelf with objects, a chandelier and a vase of flowers.

Adjust#2: A Caucasian woman with a cheerful facial expression. She is wearing red lipstick, earrings, headphones and she is wearing a red collared blouse. The woman has straight brown hair and green eyes. At the back there is a shelf with objects and books on the gray wall, a chandelier and a vase of flowers.

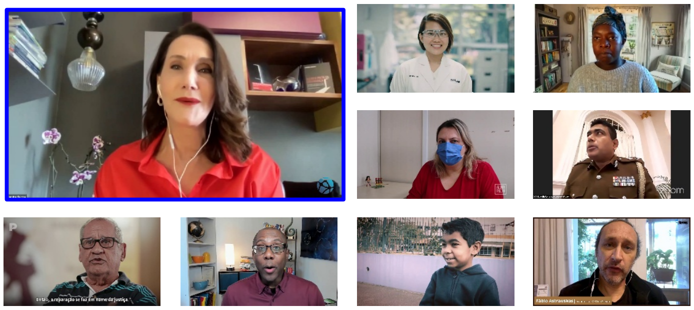

Fig. 2: Dataset samples showing data diversity and representativeness, and the set of labels of the image highlighted in blue.

measure returns the mean value among all pixels of the gray version of the image. As a high-feature of the image, we count the number of objects detected by YOLO [14].

B. Diversity and representativeness

In order to evaluate the data diversity and representativeness, we analyze the 7 features bellow. For each feature, we employed pre-trained classification and VQA models.

a) Gender: We employed the pre-trained models DeepFace [15], FairFace [16], and DEX [17], that classify an image as *female* or *male*. We also applied the VQA model BLIP [18] with the query *“What is the person’s gender?”*.

b) Age: We arrange the possible ages in 5 classes recommended by the World Health Organization [19]: ≤ 9 for *children*, $10 - 19$ for *youth*, $20 - 44$ for *adults*, $45 - 64$ for *middle-aged* individuals, and ≥ 65 for the *elderly*. The same models used for gender classification were applied to estimate age. For BLIP, we input the query *“How old is the person?”*.

c) Accessories: Considering the diverse range of accessory options, we focused on 9 specific classes based on their frequency in this domain and the union of the applied models outputs: *glasses/sunglasses*, *headphones/headsets*, *jewelry* (including necklaces, earrings, rings, *etc.*), *hats*, *scarves*, *masks*, *microphones*, *turban*, and an *others* category for cases in which none of the other classes applies. We used the BLIP with the binary query: *“Is this person wearing $ACCESSORY_CLASS$?”*. Additionally, we utilized DenseCap [13] pre-trained model to generate dense captions of the image. We filtered only the captions with positive confidence scores. For each word in the captions, we verify if it is associated with any classes (excluding the “others” class) by checking if they are synonyms, hyponyms, or exact matches of those classes using the sets of cognitive synonyms from WordNet [20].

d) Ethnicity: We included 6 classes based on the intersection of outputs of the applied models: *Indian*, *Middle Eastern*, *Black*, *Latino*, *White*, and *Asian*. We employed DeepFace and FairFace. For the latter, we considered the output *Southeast Asian* or *East Asian* as *Asian*. We also used BLIP with the query *“Is this person $ETHNICITY$?”*, in the order above, accepting the first positive answer. Additionally, we asked an open question to BLIP, *“What is the person’s ethnicity?”*. The reason is twofold: to have an answer in cases that the answers to all six binary questions were negative, and to tiebreak scenarios in which the answers to the binary questions related to *Middle Eastern* and *Indian* were positive.

e) Facial Emotion: We selected 7 classes from the intersection of the outputs of the models: *angry*, *fearful*, *neutral*, *sad*, *disgusted*, *happy*, and *surprised*. We employed DeepFace, Emotion-detection [21], FER [22], and PAZ [23]. BLIP was used with the query *“Between angry, fear, neutral, sad, disgust, happy or surprise, what is the person’s emotion?”*.

f) Scene: Scene classification involved 9 classes based on their relevance and frequency in this domain: *living room*, *office*, *library*, *kitchen*, *studio* (including music studio, art studio, meeting studio, etc.), *bedroom*, *outside*, and *others* for images that did not fit into these specific scenes. For this task, we employed the pre-trained model PlacesCNN [24] with 365 scene classes and a PlacesCNN model that we fine-tuned in only 19 scene classes more related to webinar videos. We also used BLIP with the query *“What room is the person in?”*.

g) Clothes: Due to the majority of images showing people from the waist up, clothes classification models could potentially lead to confusion between classes, *e.g.*, a shirt and a dress. Therefore, we aggregate cloth classes into 3 categories: *informal*, *formal*, and *uniform*. We employed DenseCap in conjunction with WordNet to generate dense captions and identify words related to clothing. We also applied BLIP with the query *“What is the person wearing?”* and manually categorized into these 3 classes. We designated answers such as *suit*, *tie*, *vest*, and *blazer* as formal, while responses including *uniform*, *lab coat*, and *nurse uniform* were classified as uniform. Any other responses were considered informal.

C. Implementation Details

In the process of creating the dataset, we employed the MTCNN [25] for face detection and the Python package Face Recognition[2] for face descriptor. To translate the annotated descriptions, we used the Deep-Translator[3]. Regarding the constants, we empirically set $\Delta = 120$, $\gamma = 5$, $\alpha = 0.11$, and $\beta = 0.6$. We replaced the Haar Cascade Face classifier used on the original implementation of DEX, Emotion-detection, and PAZ by the MTCNN model for the sake of fairness.

[2] Publicly available at https://github.com/ageitgey/face_recognition

[3] Publicly available at https://github.com/nidhaloff/deep-translator

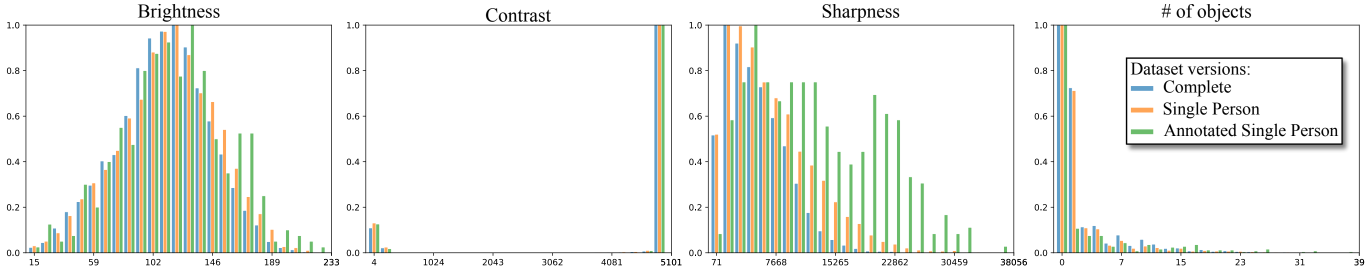

Fig. 3: Normalized histograms for the image quality analysis depicting low and high-level features of the images.

D. Results and Discussion

From the results related to the image quality analysis presented in Fig. 3, we can observe that, regarding brightness, only a few images were either excessively bright or dim. In terms of contrast, most images have high contrast, indicating images with a wide range of tones. In the sharpness analysis, the majority of images returned a good level of details. Most of the images depict scenes that are not densely populated with objects. Most importantly, all three dataset versions exhibit similar distributions over these aspects, suggesting that the smaller versions are representative samples of the larger ones.

We have included the diversity and representativeness analysis of our datasets in Tab. I, where the "Coverage" column indicates the percentage of data with a valid output for each method. Sample mean and standard deviation are calculated across dataset versions, aiming to evaluate the representativeness of the smaller versions related to the complete dataset. Therefore, we aimed for a lower standard deviation value.

Regarding gender analysis (Tab. Ia), all methods consistently exhibited a pattern of mean and standard deviation, except for DeepFace, which displayed a significant bias towards the male class. When analyzing clothing (Tab. Ib), we noticed that all models recognized more informal than formal clothes, and only BLIP recognized the uniform class. Concerning age (Tab. Ic), we observed that DeepFace had a strong tendency to output adults, despite the presence of children, young and elderly individuals in our images. FairFace is the only model with a training strategy to mitigate bias [16], and its results are a showcase of our commitment to diversity and representativeness of the data. For accessories (Tab. If), DenseCap has a known bias with classes such as glasses, hat, jewelry, and scarf, while does not recognize some of the other classes. BLIP achieved a more balanced distribution of classes, due to its training set being more representative.

As depicted in Tab. Ie, race is not a consensus among the models. Most of the "Asian" outputs by the FairFace are miss classifications due to individuals with closed eyes. DeepFace predicted "White" half of the images, whereas the remaining predictions were evenly distributed among the other classes, except for "Indian". Unlike the other models, BLIP results were more evenly distributed, except for the "Asian". We noticed that original PlaceCNN exhibited an excessive number of "other" class due to its wide range of scene possibilities. BLIP achieved more diversified results.

Finally, we highlight that only DenseCap and BLIP did not achieve 100% coverage in one task each, demonstrating the quality of the selected images. Low values of standard deviation indicates that both smaller dataset versions are reliable representations of the complete dataset.

Failure cases: Setting a cosine distance threshold for dataset construction proved challenging. While a low threshold includes multiple images of the same person, a high value does not cover all individuals in a video. In rare cases, we faced multiple images of the same person for the same video.

V. Conclusion

This work proposed an automatic data-collecting approach to create a dataset of images with foreground presenters, such as webinars, talk shows, and news. Aiming to enable accessibility for visually impaired people, we also proposed an annotation protocol based on rules specific to that audience. We demonstrated the viability of the proposed solutions by creating a dataset available in three versions: the complete dataset, a version with unique appearances of individuals, and another version featuring labeled images. Through extensive evaluation protocol, we demonstrated the quality, diversity, and representativeness of the data that comprise the datasets.

Acknowledgment

The authors would like to thank CAPES, FAPEMIG, PIBIC-UFV/CNPq, CNPq, and BICJr-UFV/FAPEMIG for funding different parts of this work; Ricson Vilaça for fine-tunning the scene classification CNN; and our tireless annotators Allan Lopes, Júlia Vieira, Júlia Lopes, and Sophia Jorge.

References

[1] B. Nascimento, M. Ribeiro, L. Silva, N. Capobiango, and M. Silva, "A soybean seedlings dataset for soil condition and genotype classification," in SIBGRAPI, vol. 1, 2022, pp. 85–90.

[2] W. Tang, D.-e. Liu, X. Zhao, Z. Chen, and C. Zhao, "A dataset for the recognition of obstacles on blind sidewalk," UAIS, vol. 22, no. 1, pp. 69–82, 2023.

[3] D. Massiceti, L. Zintgraf, J. Bronskill, L. Theodorou, M. T. Harris, E. Cutrell, C. Morrison, K. Hofmann, and S. Stumpf, "Orbit: A real-world few-shot dataset for teachable object recognition," in ICCV, 2021, pp. 10 798–10 808.

[4] C. A. G. Passamani, V. N. Neves, L. J. L. Júnior, T. Oliveira-Santos, C. Badue, and A. F. De Souza, "A method to estimate covid-19 contamination risk based on social distancing and face mask detection using convolutional neural networks," in SIBGRAPI, vol. 1, 2022, pp. 282–287.

TABLE I: Diversity and representativeness analysis of the data compounding the dataset.

Model	**Female**	**Male**	**Coverage**
DEX	46.0 ± 2.8	54.0 ± 2.8	100.0 ± 0.0
DeepFace	22.4 ± 2.4	77.6 ± 2.4	100.0 ± 0.0
BLIP	40.2 ± 6.9	59.8 ± 6.9	100.0 ± 0.0
FairFace	46.0 ± 5.5	54.0 ± 5.5	100.0 ± 0.0

(a) Gender analysis.

Model	**Informal**	**Formal**	**Uniforme**	**Coverage**
BLIP	83.9 ± 1.5	15.7 ± 1.3	0.2 ± 0.2	100.0 ± 0.0
Densecap	63.6 ± 2.5	13.5 ± 1.5	0.0 ± 0.0	75.5 ± 1.3

(b) Clothes type analysis.

Model	**Child**	**Young**	**Adult**	**Middle-aged**	**Elderly**	**Coverage**
DEX	0.0 ± 0.0	0.4 ± 0.1	72.1 ± 1.1	26.9 ± 0.9	0.6 ± 0.3	100.0 ± 0.0
DeepFace	0.0 ± 0.0	0.0 ± 0.0	95.3 ± 0.5	4.7 ± 0.5	0.0 ± 0.0	100.0 ± 0.0
BLIP	0.2 ± 0.2	1.6 ± 0.3	67.6 ± 1.3	26.3 ± 0.6	4.3 ± 0.4	100.0 ± 0.0
FairFace	0.7 ± 0.0	5.8 ± 0.9	90.6 ± 1.2	2.4 ± 0.5	0.4 ± 0.2	100.0 ± 0.0

(c) Analysis related to the age ranges defined by the World Health Organization.

Model	**Angry**	**Fearful**	**Neutral**	**Sad**	**Disgusted**	**Happy**	**Surprise**	**Coverage**
PAZ	8.9 ± 0.8	5.0 ± 1.2	31.4 ± 3.1	27.2 ± 0.6	0.0 ± 0.0	22.7 ± 1.7	4.7 ± 0.2	100.0 ± 0.0
DeepFace	9.7 ± 0.2	14.6 ± 0.4	34.3 ± 1.0	23.3 ± 2.5	0.9 ± 0.3	14.4 ± 3.8	2.8 ± 0.2	100.0 ± 0.0
BLIP	2.8 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	67.0 ± 6.0	0.0 ± 0.0	30.1 ± 7.3	0.0 ± 0.0	100.0 ± 0.0
Emotion	11.5 ± 0.3	20.4 ± 0.3	10.0 ± 0.2	35.7 ± 0.9	0.1 ± 0.1	19.5 ± 1.2	2.8 ± 0.3	100.0 ± 0.0
FER	16.0 ± 1.9	6.5 ± 0.6	42.0 ± 2.2	18.0 ± 0.5	0.0 ± 0.0	9.3 ± 1.8	8.2 ± 0.5	100.0 ± 0.0

(d) Analysis of the presenter's face emotion (inferred entirely based on visual information).

Model	**Asian**	**White**	**Middle eastern**	**Indian**	**Latino**	**Black**	**Coverage**
DeepFace	11.2 ± 1.1	50.2 ± 2.1	11.5 ± 2.7	1.6 ± 0.5	14.8 ± 3.3	10.8 ± 3.6	100.0 ± 0.0
BLIP	1.3 ± 0.1	33.5 ± 1.6	12.9 ± 1.6	5.0 ± 1.5	30.2 ± 4.2	16.8 ± 8.2	99.7 ± 0.1
FairFace	57.4 ± 2.9	21.3 ± 0.3	7.7 ± 1.3	0.9 ± 0.2	5.5 ± 1.1	7.2 ± 2.6	100.0 ± 0.0

(e) Analysis of the presenter's race (inferred entirely based on visual information).

Model	**Glasses**	**Headphones**	**Jewerly**	**Hat**	**Scarf**	**Mask**	**Microphone**	**Turban**	**Others**	**Coverage**
BLIP	26.2 ± 1.1	17.5 ± 3.4	23.2 ± 3.2	2.0 ± 1.1	2.1 ± 0.5	0.5 ± 0.4	7.7 ± 0.8	1.0 ± 0.2	19.8 ± 2.0	100.0 ± 0.0
DenseCap	50.1 ± 2.9	1.2 ± 2.1	4.2 ± 1.5	6.2 ± 2.5	0.5 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	37.9 ± 1.7	100.0 ± 0.0

(f) Analysis of the accessories that are being worn by the person.

Model	**Living Room**	**Office**	**Library**	**Kitchen**	**Studio**	**Bedroom**	**Hospital**	**Outside**	**Others**	**Coverage**
PlacesCNN	0.1 ± 0.1	9.0 ± 2.0	0.0 ± 0.0	0.1 ± 0.1	2.5 ± 0.5	0.0 ± 0.0	0.3 ± 0.1	0.0 ± 0.0	88.1 ± 1.7	100.0 ± 0.0
PlacesCNN-Finetune	0.2 ± 0.1	1, 5 ± 0.4	3.3 ± 1.1	3.8 ± 1.3	5.7 ± 1.9	73.0 ± 1.4	0.1 ± 0.1	0.5 ± 0.4	12.0 ± 1.0	100.0 ± 0.0
BLIP	42.3 ± 3.3	29.0 ± 0.7	8.5 ± 4.0	3.3 ± 0.8	0.1 ± 0.1	2.3 ± 0.3	0.1 ± 0.0	2.2 ± 1.5	12.3 ± 4.4	100.0 ± 0.0

(g) Analysis of the presenter's surroundings.

[5] H. Shah, R. Shah, S. Shah, and P. Sharma, "Dangerous object detection for visually impaired people using computer vision," in AIMV, 2021, pp. 1–6.

[6] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in ECCV, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 417–434.

[7] D. Gurari, Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl, and J. P. Bigham, "Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people," in CVPR, 2019, pp. 939–948.

[8] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in CVPR, 2018, pp. 3608–3617.

[9] D. L. Fernandes, M. H. F. Ribeiro, F. R. Cerqueira, and M. M. Silva, "Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs," in VISIGRAPP (5: VISAPP), 2022, pp. 526–534.

[10] ABNT, "Accessibility in communication: Audio description - NBR 16452," ABNT, 2016.

[11] V. Lewis, "How to write alt text and image descriptions for the visually impaired," Perkins S for the Blind, 2018.

[12] D. M. Kuhn and V. P. Moreira, "Brcars: a dataset for fine-grained classification of car images," in SIBGRAPI, 2021, pp. 231–238.

[13] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in CVPR, 2016, pp. 4565–4574.

[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv, 2018.

[15] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in ASYU, 2020, pp. 23–27.

[16] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in WACV, 2021, pp. 1548–1558.

[17] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deep expectation of apparent age from a single image," in ICCV, 2015, pp. 10–15.

[18] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in ICML, 2022, pp. 12 888–12 900.

[19] O. B. Ahmad, C. Boschi-Pinto, A. D. Lopez, C. J. Murray, R. Lozano, and M. Inoue, "Age standardization of rates: a new who standard," Geneva: World Health Organization, vol. 9, no. 10, pp. 1–14, 2001.

[20] G. A. Miller, "Wordnet: a lexical database for english," Comms. of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[21] I. J. Goodfellow and et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[22] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," arXiv, 2017.

[23] O. Arriaga, M. Valdenegro-Toro, M. Muthuraja, S. Devaramani, and F. Kirchner, "Perception for autonomous systems (paz)," arXiv, 2020.

[24] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," TPAMI, 2017.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.