Let Me Take a Better Look: Towards Video-Based Age Estimation

Krešimir Bešenić[1][a], Igor S. Pandžić[1][b] and Jörgen Ahlberg[2][c]
[1]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia
[2]Computer Vision Laboratory, Linköping University, 58183 Linköping, Sweden

Keywords: Age, Video, Benchmark, Semi-Supervised, Pseudo-Labeling.

Abstract: Taking a better look at subjects of interest helps humans to improve confidence in their age estimation. Unlike still images, sequences offer spatio-temporal dynamic information that contains many cues related to age progression. A review of previous work on video-based age estimation indicates that this is an underexplored field of research. This may be caused by a lack of well-defined and publicly accessible video benchmark protocol, as well as the absence of video-oriented training data. To address the former issue, we propose a carefully designed video age estimation benchmark protocol and make it publicly available. To address the latter issue, we design a video-specific age estimation method that leverages pseudo-labeling and semi-supervised learning. Our results show that the proposed method outperforms image-based baselines on both offline and online benchmark protocols, while the online estimation stability is improved by more than 50%.

1 INTRODUCTION

Human visual and cognitive systems allow us to perform many incredibly complex tasks effortlessly. However, estimating biological age from unknown faces based only on visual cues is a challenging task, even for humans. Research suggests that humans misestimate age from facial images by 4.7 to 7.2 years on average (Han et al., 2013). Large variations in apparent age for subjects of the same biological age can be caused not only by genetic predispositions and health, but also by many different external factors such as living conditions, weather exposure, facial cosmetics, surgical operations, facial hair, and even facial expressions (Dibeklioğlu et al., 2015).

While automatic age estimation has prominent application fields, such as human-computer interaction (HCI), precision advertising, and the beauty industry, some more strict fields, such as security, forensics, and law enforcement, are yet to reach widespread use. Age estimation models often fail under in-the-wild conditions by a margin that is not acceptable for such demanding application fields.

Humans and ML models can estimate biological age from a single image with comparable accuracy. However, when humans are not confident in their estimation, they tend to *take a better look* by examining the subject for a longer period of time and from different viewpoints. This observation is supported by the review in (Hadid, 2011) stating that psychological and neural studies (Bassili, 1979; Hill and Johnston, 2001; Knight and Johnston, 1997; O'Toole et al., 2002) indicate that head-pose and facial expression changes provide important cues for face analysis. Existing image-based methods can be applied directly to video frames. However, (Ji et al., 2018) point out that deploying image-based age estimation models directly to videos leads to estimation stability issues, while (Hadid, 2011) states that the image-based approach exploits the abundance of frames in videos but ignores useful temporal correlations and facial dynamics.

This work explores the potential of videos and video-based methods for refinement of automatic facial age estimation. Our main contributions are summarized as follows: *(i)* We review previous research and public datasets for video-based age estimation to pinpoint the main obstacles in this underresearched field. *(ii)* We reproduce the recent frame-based age estimation benchmark from (Hazirbas et al., 2021), achieve state-of-the-art result on their protocol, and make the missing benchmark metadata publicly available. *(iii)* We design a new video-based age estimation benchmark and ensure reproducibility by making the framework and metadata publicly available. *(iv)* We design a semi-supervised video age estima-

[a] https://orcid.org/0000-0002-5861-7076
[b] https://orcid.org/0000-0002-2075-8295
[c] https://orcid.org/0000-0002-6763-5487

Bešenić, K., Pandžić, I. and Ahlberg, J.
Let Me Take a Better Look: Towards Video-Based Age Estimation.
DOI: 10.5220/0012376800003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 57-69
ISBN: 978-989-758-684-2; ISSN: 2184-4313

tion method that overcomes the lack of labeled training data and outperforms its image-based counterpart, thus setting baseline results on the proposed benchmark.

2 RELATED WORK

The main premise of this study is that videos can provide extensive information useful for age estimation. However, there is still only a handful of published datasets and methods that leverage video information to improve age estimation. Systematic surveys on image-based age estimation methods and datasets can be found in (Panis et al., 2016; Angulu et al., 2018; Al-Shannaq and Elrefaei, 2019; Othmani et al., 2020; Agbo-Ajala and Viriri, 2021). This section briefly reviews video-based age estimation methods and video datasets with age annotations.

2.1 Video-Based Age Estimation Methods

According to (Hadid, 2011), there are two main strategies for video-based face analysis. The simplest strategy is to apply image-based methods to all video frames or a set of sampled frames from a video. Individual frame results are then fused across the sequence. A more elaborate strategy consists of leveraging both face appearance and face dynamics information through spatio-temporal modeling. Hadid implemented two baseline methods based on SVM classifiers, LBP features (Ojala et al., 1996; Ojala et al., 2002) for static images, and Volume-LBP features (Zhao and Pietikainen, 2007) for video sequences. They used 2,000 web-scraped videos manually annotated with apparent age labels. Their experiments indicated that the video-based approach can improve performance for face recognition, gender classification, and ethnicity classification tasks, but not for age estimation.

(Dibeklioğlu et al., 2012a) focused on the discriminative power of smile dynamics for age estimation. They leveraged movement features of facial key points to complement appearance-based LBP features and improve SVM estimation accuracy. Their experiments on the UvA-NEMO Smile Database (Dibeklioğlu et al., 2012b) showed significant performance improvement over the image-based method, as well as compared to the method from (Hadid, 2011). In (Dibeklioğlu et al., 2015), they also considered IEF (Alnajar et al., 2012), GEF (Alnajar et al., 2012), and BIF (Guo et al., 2009) features. Surface area features based on a mesh model were used instead of facial key-point movement features, followed by a novel two-level classifier. Experimentation was extended to the introduced UvA-NEMO Disgust dataset. Video-based methods significantly outperformed the image-based baseline on both versions of the UvA-NEMO dataset.

Instead of handcrafted features used in (Hadid, 2011; Dibeklioğlu et al., 2012a; Dibeklioğlu et al., 2015), (Pei et al., 2019) used a combination of CNN, RNN, and attention modules. Their proposed Spatially-Indexed Attention Model (SIAM) used a CNN for appearance modeling, a spatial attention module for the detection of salient facial regions, an RNN model for capturing facial dynamics, and a temporal attention module for temporal saliency, trained in an end-to-end manner. UvA-NEMO Smile and UvA-NEMO Disgust datasets were once again used, following protocol from (Dibeklioğlu et al., 2015). The proposed neural network-based approach outperformed previous methods based on handcrafted features. (Ji et al., 2018) pointed out that deploying image-based age estimation models directly to videos often suffers from estimation stability issues. To address this, they proposed a combination of CNN-based feature extraction and attention-based feature aggregation modules, sharing some similarities with (Pei et al., 2019). Their loss function combined MSE loss and a component for estimation stability. The model was not trained in an end-to-end manner, as the CNN was trained on the MORPH-II (Ricanek and Tesafaye, 2006) image dataset. To train the feature aggregation module, they built a video dataset comprising 18,282 frames from a single twelve-minute video of one subject. Their experiments demonstrated improvements with respect to both age estimation accuracy and stability.

To attenuate the effect of head-pose on video-based age estimation, (Han et al., 2021) based their method on pose-invariant *uv* texture maps reconstructed from video frames by a Wasserstein-based GAN. They introduced the UvAge video dataset and used it both for training and evaluation. (Zhang and Bao, 2022) combined multi-loss CNN for head-pose estimation (Ruiz et al., 2018) and DRF (Shen et al., 2018) for age estimation. To mitigate the negative effect of head-pose variation, they trained a head-pose model on the 300W-LP dataset (Zhu et al., 2016) and used it to create multiple subsets of the CACD (Chen et al., 2015) and AFAD (Niu et al., 2016) datasets. Two 12-minute facial videos from two subjects were collected for evaluation purposes. Using frontal models and selecting frames with near-frontal faces improved both age estimation accuracy and stability. Both this work and the work from (Han et al., 2021)

Table 1: Overview of the video-based age estimation datasets.

Dataset	Subjects	Videos	Age range	Demographics	Head pose	Illumination
UvA-NEMO Smile (Dibeklioğlu et al., 2012b)	400	1,240	8 - 76	unbalanced	mostly frontal	constrained
UvA-NEMO Disgust (Dibeklioğlu et al., 2015)	324	518	8 - 76	unbalanced	mostly frontal	constrained
UvAge (Han et al., 2021)	516	6,898	16 - 83	unconstrained	unconstrained	unconstrained
Casual Conversations (Hazirbas et al., 2021)	3,011	45,186	18 - 85	semi-balanced	unconstrained	unconstrained
Casual Conversations Mini (Hazirbas et al., 2021)	3,011	6,022	18 - 85	semi-balanced	unconstrained	balanced
Casual Conversations v2 (Porgali et al., 2023)	5,567	26,467	18 - 81	semi-balanced	unconstrained	unconstrained

were focused on achieving robustness to head pose changes without exploiting any temporal information from videos.

A recurring topic in the reviewed work is the lack of publicly available video data, as almost all authors resorted to data collection. However, the collected data was not made publicly available in (Hadid, 2011; Ji et al., 2018; Zhang and Bao, 2022). Manual data collection often results in very small datasets. Ji *et al.* and Zhang *et al.* used very limited amounts of video data (*i.e.*, one or two videos), undermining the reliability of their conclusions. Dibeklioğlu *et al.* based their work on a larger video dataset, but used it for both training and evaluation. A very specific nature of the used data, where every subject transitions from neutral to smiling facial expression, potentially caused overfitting to this specific type of data. They introduced a dataset related to a different facial expression (disgust), but did not perform cross-domain testing, same as Pei *et al.*. Ji *et al.*, Han *et al.*, and Zhang *et al.* did not perform a comparison with previous video-based age estimation methods.

2.2 Video-Based Age Estimation Datasets

A lack of a large public in-the-wild video-based age estimation benchmark undermines the convincingness of some of the reviewed findings and the ability to fairly compare introduced methods. This section reviews video-based age estimation datasets used in the reviewed work, as well as some recent datasets suitable for this purpose.

UvA-NEMO Smile Dataset was initially introduced to study differences between spontaneous and posed smiles in (Dibeklioğlu et al., 2012b). The dataset consists of 597 spontaneous and 643 posed smile recordings, totaling in 1,240 videos. The videos were collected from 400 volunteers (185 female and 215 male) with ages ranging from 8 to 76 years. Subjects were mostly Caucasian. The recordings were done in a controlled environment, with constrained illumination and high-resolution cameras. Each video segment starts with a neutral expression and transitions to a smiling expression.

UvA-NEMO Disgust Dataset (Dibeklioğlu et al., 2015) was collected concurrently with the UvA-NEMO Smile Dataset following the same recording setup. 324 volunteers (152 female, 172 male) were recorded posing disgust facial expressions. 313 of them also participated in the UvA-NEMO Smile Dataset collection. Similar to the Smile version of the dataset, age varies from 8 to 76 years, and the subjects are mostly Caucasian. Each of the 518 disgust video segments once again starts with a neutral expression and transitions to the target expression.

UvAge Dataset (Han et al., 2021) was created specifically for age estimation from videos. It consists of 6,898 videos from 516 subjects. The proposed web-scraping technique was based on collection of videos of celebrities with birth information available on Wikipedia. To get a reliable video recording time, they used traceable public events such as the Academy Awards or the G20 summit. The videos were manually verified and segmented into sequences containing only one subject. Along with age labels, each video was also annotated with identity, gender, ethnicity, and occupation.

Casual Conversations Dataset (CC) (Hazirbas et al., 2021) is a recently published video dataset from Facebook (Meta) AI designed for measuring fairness of computer vision and audio models across a diverse set of ages, genders, apparent skin tones, and ambient lighting conditions. It consists of 45,186 high-quality videos collected from 3,011 paid individuals who agreed to participate in the project and explicitly provided age and gender labels themselves. A group of trained annotators additionally labeled skin tone according to Fitzpatrick scale (Fitzpatrick, 1975) and ambient light type. The authors also proposed a well-balanced subset of the CC dataset, denoted as Casual Conversations Mini (CCMini). The subset was formed by selecting one dark and one bright video per subject (when possible) to have a balanced lighting distribution, with a total of 6,022 videos.

Casual Conversations v2 Dataset (Porgali et al., 2023) is a follow-up to the original CC dataset (CCv2), curated with special focus on geographical diversity. 5,567 subjects from 7 countries were paid to participate in data collection. The 26,467 collected videos amount to 320 hours of scripted and 354 hours of non-scripted conversations. Some of the labels, such as age, gender, language/dialect, disability, physical adornments/attributes, and geo-location, were self-provided by the participants. Trained annotators were used to additionally label for Fitzpatrick Skin Type (Fitzpatrick, 1975), Monk Skin Tone (Monk, 2014), voice timbre, recording setup, and per-second activity.

A summary of the reviewed datasets is presented in Table 1. Our efforts to receive access to UvA-NEMO Smile, UvA-NEMO Disgust, and UvAge datasets were unsuccessful, making the Casual Conversation datasets the only publicly accessible resource suitable for video-based age estimation research. However, the CC dataset licenses permit the use of provided labels only for evaluation purposes.

3 VIDEO AGE ESTIMATION BENCHMARK DATA

Not only that Casual Conversations are the only publicly accessible video-based datasets, they were also curated with a special focus on ethical data collection and demographic fairness. Moreover, they are the largest video datasets with precise self-provided age annotations, and their licenses allow for evaluation of both academic and commercial models[1]. All this makes them great candidates for a video age estimation benchmark.

3.1 CCMiniIMG

The authors of the CC dataset explored its potential for age estimation evaluation on CCMini; a well-balanced subset of the CC dataset. Their benchmark data consists of 100 frames sampled from each of the videos.

For further reference, we dub this image set as CCMiniIMG. Although this initial effort toward a CC-based age estimation benchmark motivated our work, we discuss its adequacy for video-based age estimation. CCMiniIMG is made available in the form of pre-extracted raw video frames, where 100 frames are provided for each of 6,022 videos. However, the authors did not provide face detection metadata or clear information on the frame sampling procedure. The lack of face detection metadata prevents the exact reproduction of their evaluation protocol as some frames contain multiple subjects. The lack of information on the frame sampling method and the fact that the protocol relies only on frames rather than continuous sequences prevents the evaluation of video-based age estimation methods that leverage temporal information.

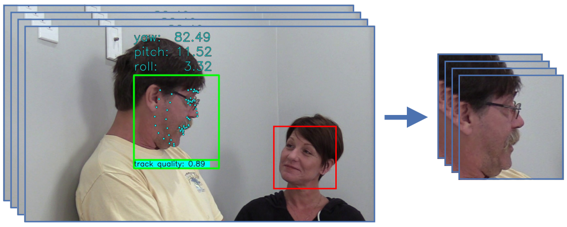

Figure 1: CCMiniVID video processing framework. All faces are tracked with a commercial tracking system. The metadata of multi-subject videos is manually filtered to contain only data related to the subject of interest. The framework extracts aligned face crop sequences using the raw CC videos and the produced tracking data.

3.2 CCMiniVID

To design a video benchmark dataset for age estimation based on the CC data, we follow CCMiniIMG and select 6,022 well-balanced videos from the CCMini subset. All videos were recorded at 30 FPS and the mean video duration is 64.44 ± 13.56 seconds, with 99% of videos lasting for at least 20 seconds. We set our target to extract sequences with 20 seconds of continuous face presence from each video, where possible. We utilize a commercial face tracking system from Visage Technologies[2] to produce high-quality frame-level tracking data comprising 75 facial landmark points, apparent pitch, yaw, and roll head angles, tracking quality, and face scale. We proceed with semi-automatic tracking data analysis, as shown in Figure 1.

The CC data does not provide correspondence between subjects and labels for the multi-subject videos, potentially causing erroneous subject-level labels. We perform automatic detection of multi-subject video candidates, followed by manual verification and selection of tracking streams in 108 videos. In two of the videos subjects of interest are facing away from the camera, making their faces fully self-occluded. These videos are not suitable for a face-oriented evaluation benchmark. Continuous face tracking with a target duration of 20 seconds was achieved for 5,932 of 6,022 videos. Manual verification of the remaining 90 videos showed that in 21 videos continuous tracking was not sustainable due to occlusions, self-occlusions, camera issues, or extreme lighting. 69 videos were shorter than 20 seconds, with the shortest one lasting for only 6 seconds.

[1] For specific conditions please refer to the CC license agreements.

[2] https://visagetechnologies.com/facetrack/

Table 2: Age classification accuracy results for the baseline methods from (Hazirbas et al., 2021) and our image-based age estimation model from (Bešenić et al., 2022) on the CCMiniIMG benchmark data according to the image-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
(Levi and Hassner, 2015)	38.05	37.44	39.48	66.67	39.56	38.72	40.84	36.47	36.47	34.89	38.49	37.04
(Lee et al., 2018)	42.26	42.28	44.53	100.00	42.33	41.78	42.30	42.79	42.44	37.99	42.94	41.12
(Serengil and Ozpinar, 2020)	54.32	54.21	56.18	83.33	46.51	55.52	54.59	55.78	53.78	52.57	54.17	55.20
(Bešenić et al., 2022)	73.06	70.20	76.48	87.50	76.72	82.15	72.90	70.24	67.52	65.61	73.90	71.64

Table 3: Age classification accuracy results for our image-based age estimation model from (Bešenić et al., 2022) on the three versions of CCMiniVID benchmark data according to the image-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniIMG	73.06	70.20	76.48	87.50	76.72	82.15	72.90	70.24	67.52	65.61	73.90	71.64
CCMiniVID-O	73.60	70.90	76.75	93.75	77.59	81.31	74.25	74.09	65.82	67.48	75.18	70.96
CCMiniVID-A	73.50	70.56	77.02	87.50	77.59	81.55	73.13	74.09	66.67	67.24	75.00	70.98
CCMiniVID-R	73.59	70.85	76.81	93.55	77.23	81.38	74.43	73.88	65.70	67.49	75.17	70.99

Based on the obtained tracking data and manual verification, we propose three versions of the CCMiniVID benchmark data. CCMiniVID-O, where O stands for ”original”, comprises only videos from the original CCMini subset. It contains 5,932 sequences with continuously tracked 600 frames (20s@30FPS) and 90 outlier videos that have between 100 and 599 continuously tracked frames. It also contains the aforementioned 2 videos where subjects of interest are not visible. In CCMiniVID-A, where A stands for ”alternatives”, replacements for 92 problematic videos were manually selected from the full CC dataset by looking for the most similar substitutes in video galleries of target subjects. This allows for consistent evaluation with 600 consecutively tracked frames for all 3,011 subjects from the CC dataset. Additionally, we propose CCMiniVID-R, where R stands for ”reduced”; a simple subset of CCMiniVID-O where the 92 problematic videos were removed. This allows for precise and consistent testing without the need for downloading the full 6.9 TB CC dataset. Figure 2 presents distributions for CCMiniVID-R main labels, while a broader explorative analysis is available in Appendix B.

4 VIDEO AGE ESTIMATION BENCHMARK PROTOCOL

To design the benchmark protocol, we first review the image-based age classification protocol introduced in (Hazirbas et al., 2021).

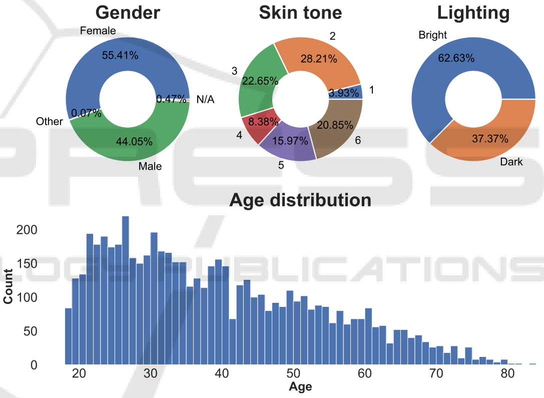

Figure 2: CCMiniVID-R label distributions for gender, age, and skin tone (subject-level) and lighting (video-level).

4.1 Image-Based Benchmark Protocol

The authors of (Hazirbas et al., 2021) extracted 100 frames from each of the 6,022 CCMiniIMG videos and detected faces with the DLIB face detector (King, 2009). The authors provide only video-level metadata and raw frames. Face detection metadata was not provided, and neither face detection setup nor frame sampling algorithm were specified. In the proposed protocol, age estimation is treated as a 3-class age classification task. The median of 100 image-level age estimations is used as the video-level estimation which is then mapped to 3 predefined age groups (i.e., 18-30, 31-45, and 46-85). Classification accuracy was selected as the main metric. The protocol additionally relies on the dataset’s auxiliary labels to calculate accuracy across 3 gender groups, 6 apparent skin tone types, and 2 ambient lighting types. The protocol reproduction details and metadata are available in Appendix A.

4.1.1 Image-Based Baseline Results

We evaluate the performance of our best-performing image-based age estimation model from (Bešenić et al., 2022) and compare it to the previously reported SOTA results from (Hazirbas et al., 2021). The selected model is based on a CNN architecture from MobileFaceNet family (Chen et al., 2018) and trained on the large in-the-wild B3FD image dataset (Bešenić et al., 2022) using the DLDLv2 age estimation method (Gao et al., 2018). The model takes 128×128 RGB face crop inputs and outputs 101 values corresponding to ages from 0 to 100. More details are available in (Bešenić et al., 2022).

Table 2 presents the baseline results. The selected model outperforms the overall age classification accuracy of the previously reported leading method (Serengil and Ozpinar, 2020) by a large margin of 18.74 points, establishing the new state-of-the-art on this benchmark. Although it outperforms the baseline methods by a large margin, it shows similar relative performance drop for female subjects, darker skin tones, and poorly illuminated recordings.

To validate the CCMiniVID benchmark data from Section 3, we apply the image-level baseline model and protocol to its three versions. Table 3 shows a comparison with results obtained on the original CCMiniIMG data. The results are closely matched, showing that while the CCMiniVID benchmark data offers several benefits over CCMiniIMG, such as continuous video sequences, face tracking data, and clear mapping between video subjects and the dataset's metadata, it retains a very similar difficulty level.

4.2 Video-Based Benchmark Protocol

Methods that work with sequential (i.e., temporal) data can be either offline or online. Offline methods process a video as a unit non-causally and produce a single estimate for the whole video. Temporal stability is not an issue since only one estimate per video is produced. Therefore, we find Mean Absolute Error (MAE) to be a sufficient metric for offline estimation methods. The absolute error is calculated based on a single estimate per video, while the mean is calculated over all videos in the dataset. Online methods produce updated age estimations in real-time as new video frames are captured. For online methods, we propose Temporal Mean Absolute Error ($tMAE$) and Temporal Standard Deviation ($tSTD$) as the benchmark metrics to evaluate both method's accuracy and online estimation stability, respectively. These are standard MAE and STD metrics calculated at the frame level, as online methods produce new estimates for each of the frames in the temporal dimension. Following the CCMiniIMG protocol, we rely on the dataset's additional labels to calculate the proposed metrics across 3 gender groups, 6 apparent skin tone types, and 2 ambient lighting types. To unambiguously define the benchmark protocol, we make the metadata (including the face tracking data from the commercial tracking system), the test set extraction framework, and the evaluation framework publicly available at https://github.com/kbesenic/CCMiniVID.

4.2.1 Video-Based Baseline Results

Tables 4 and 5 present results of our image-based age estimation model described in Section 4.1.1 on the three CCMiniVID benchmark dataset versions, following the previously described offline and online video protocols, respectively. To comply with the offline estimation protocol, which expects a single estimate per video, we use the median of the frame-level estimations to calculate MAE. The $tMAE$ and $tSTD$ metrics for the online protocol are calculated directly from the frame-level estimation errors.

The results obtained on the three versions are very similar, and the relative performance drop can be observed in the case of female subjects, darker skin tones, and lack of good lighting. High $tSTD$ numbers indicate unstable age estimation across video frames. The large gap between overall offline and online MAE indicates that video-level estimations are more precise than frame-level estimations.

5 TOWARDS VIDEO-BASED AGE ESTIMATION METHOD

All previously presented results are obtained with image-based age estimation methods. The main premise of this work is that age can be estimated more precisely by *taking a better look*, i.e. by using a longer video sequence instead of a single image. To verify this, we first analyze how video sequence duration affects the age estimation accuracy.

5.1 Taking a Better Look

For this experiment, we use the CCMiniVID-A video benchmark data which contains 600 frames (20s@30FPS) of continuous face tracking for two videos of each subject in the CCMini dataset. We find this version of the benchmark data most suitable

Table 4: Offline age estimation *MAE* for our image-based age estimation model from (Bešenić et al., 2022) on the three versions of CCMiniVID benchmark data according to the offline video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniVID-O	5.25	5.61	4.81	3.82	5.14	4.78	5.02	5.15	5.56	5.95	5.04	5.60
CCMiniVID-A	5.25	5.64	4.78	3.63	5.03	4.74	5.05	5.14	5.59	6.00	5.00	5.67
CCMiniVID-R	5.23	5.60	4.78	3.89	4.94	4.79	4.97	5.19	5.57	5.92	5.01	5.59

Table 5: Online age estimation *tMAE* and *tSTD* for our image-based age estimation model from (Bešenić et al., 2022) on the three versions of CCMiniVID benchmark data according to the online video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniVID-O	6.19±3.25	6.77±3.73	5.47±2.66	4.77±2.61	6.19±3.44	5.64±2.98	5.84±2.99	5.98±3.01	6.63±3.60	7.04±3.71	5.97±3.13	6.56±3.46
CCMiniVID-A	6.10±3.12	6.68±3.58	5.39±2.54	4.43±2.54	6.03±3.33	5.53±2.84	5.78±2.88	5.84±2.84	6.58±3.44	6.98±3.57	5.85±3.02	6.52±3.29
CCMiniVID-R	6.17±3.25	6.75±3.73	5.45±2.66	4.83±2.61	6.02±3.44	5.64±2.98	5.79±2.98	6.01±3.02	6.64±3.60	7.01±3.71	5.94±3.13	6.54±3.46

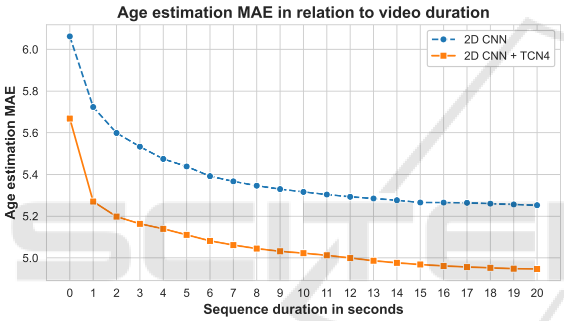

Figure 3: Age estimation *MAE* of our 2D CNN image-based age estimation model from (Bešenić et al., 2022) and the proposed video-based model (2D CNN + TCN4) in relation to video sequence duration on the CCMiniVID-A offline video-based benchmark protocol.

since it contains all subjects while all video subsequences are of the exact same duration. The majority of video-based methods rely on frame subsampling techniques to avoid redundant processing of nearly identical neighboring frames and to reduce computational complexity. We set the subsampling step to 6, following various video processing methods (Gao et al., 2017; Xu et al., 2019; Eun et al., 2021; Xu et al., 2021; Chen et al., 2022). We calculate age estimation *MAE* for video subsequences lasting from a single frame (0 seconds) to 20 seconds. The results are presented in Figure 3, denoted as 2D CNN.

By using video sequences of 1 second, the single-frame-based estimation error is reduced by 5.59%. By using 5 seconds, the error is reduced by 10.29%. Increasing sequence duration from 5 to 15 seconds reduces the error by an additional 2.67 percentage points. However, increasing the sequence duration from 15 to 20 seconds results in almost no improvement. Using the full sequences reduces the estimation error by 13.35%, compared to the single-frame approach. The results demonstrate that *taking a better look* is very useful for age estimation, even with a basic image-based estimation method and simple median aggregation of the frame-level results.

5.2 Cherry-Picking Based on Tracking Data

Cherry-picking is a popular name for a technique based on selection of the best or most suitable elements from a set. A frontal-face cherry-picking approach was proposed by (Zhang and Bao, 2022), where a fixed threshold was used on the sum of absolute head-pose angles. To evaluate the cherry-picking approach, we can utilize the produced frame-level tracking data described in Section 3.2 (*i.e.*, head-pose angles, tracking quality, and face scale). Using a fixed threshold is not suitable for our evaluation protocol since it might cause biases in video-based or even subject-based label distributions. For example, there might be more female subjects with bad lighting in the non-frontal subset. To mitigate this issue, we extract three equally sized subsets from each of the benchmark videos. As a baseline, we extract the chronologically first 50% of the video frames. We also extract the best and the worst 50% of the video frames based on a certain criterion. This ensures that all videos and all subjects are used in all three subsets and that the subsets are of equal size. Results of this experiment are presented in Table 6.

Contrary to the findings in (Zhang and Bao, 2022), head-pose angles seem to have negligible influence on age estimation error in our experiments. The used

Table 6: Age estimation *MAE* of our image-based age estimation model from (Bešenić et al., 2022) on the CCMiniVID-A benchmark data for frame cherry-picking approaches based on different face tracking data.

Criterion	Yaw	Pitch	Roll	Tracking quality	Face scale
First 50%	5.29	5.29	5.29	5.29	5.29
Best 50%	5.32	5.26	5.29	5.27	5.27
Worst 50%	5.26	5.32	5.26	5.30	5.31

image-based age estimation model was trained on the B3FD dataset (Bešenić et al., 2022), a very large in-the-wild age estimation dataset with unconstrained head poses. The model was also trained with data augmentations to make it robust to tracking instabilities and low face resolution, further explaining why filtering based on auxiliary tracking data did not result in significant improvements.

5.3 Video-Based Age Estimation Method

The main obstacle to training an age estimation model that leverages facial dynamics and temporal information from videos is the lack of video data with age labels. As mentioned in our data review in Section 2.2, there are no publicly accessible video datasets that permit training of age models. As reviewed in Section 2.1, some researchers resorted to a manual collection of very small video datasets (*e.g.*, only one or two subjects). Our opinion is that this is not enough to train a reliable model with generalization capabilities.

To deal with the lack of annotated video data that restricts usability of fully supervised learning, we explore pseudo-labeling and semi-supervised learning. Whereas supervised learning requires labels for all training samples, semi-supervised learning algorithms aim at improving their performance by utilizing unlabeled data (Oliver et al., 2018). One approach for making use of unlabeled data is generating pseudo labels. Pseudo labels are weak labels generated by the model itself and subsequently used to further train the model (Hu et al., 2021). In our setup, we leverage the fact that subject's age does not change during a single video recording. We rely on an image-based age estimation model and apply it to every frame of an unlabeled facial video dataset. The frame-level results are aggregated to get video-level pseudo labels. These pseudo labels can then be used to supervise learning of spatio-temporal models. In case of end-to-end training, the image-based backbone used to generate the pseudo-labels is also further optimized and improved.

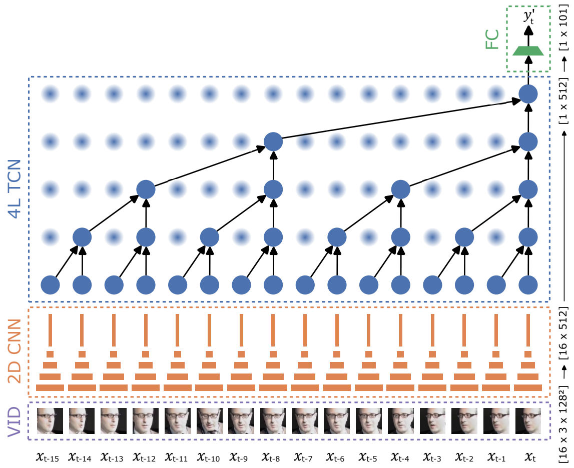

Figure 4: The proposed video age estimation method based on a 2D CNN feature extractor and a 4-layer TCN temporal model (4L TCN), followed by a fully connected layer (FC) for dimensionality reduction. The model takes 16 input video frames (VID) denoted as x_t to x_{t-15} to produce an age estimation probability vector y'_t.

5.3.1 Pseudo-Labeled Training Data

The main advantage of the proposed approach is that any source of unlabeled facial videos can be used for model training. We combine two large sources of raw videos. The video portion of the IJB-C dataset (Maze et al., 2018) is selected since the dataset's general statistics indicate good age distribution. CelebV-HQ dataset (Zhu et al., 2022) is selected for its size and good distribution with respect to facial expressions, appearance attributes, and actions. The raw videos were once again processed with the commercial face tracking system and frame-level estimations were produced with the image-based age estimation model from Section 4.1.1. The median was used to aggregate frame-level estimations into video-level pseudo labels. We filtered the video data with respect to face scale, tracking quality, and sequence duration. The produced set consists of 28,619 videos with a total of 7,535,299 frames, averaging 263 continuously tracked frames per video. Pseudo-label distribution ranges from 6 to 89 years.

5.3.2 Semi-Supervised Video-Based Method

In Section 4.2.1 we demonstrated that offline median estimation across video sequences can be much more precise than frame-level predictions. Our goal is to use the median pseudo labels to train a temporal model that will be able to replicate that performance boost, but in an online fashion and based on a much shorter time window.

Table 7: Offline age estimation *MAE* for our 2D CNN image-based age estimation model (Bešenić et al., 2022) and the proposed video-based age estimation models (2D CNN + TCN) on the CCMiniVID-A benchmark data according to the offline video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
2D CNN	5.25	5.64	4.78	3.63	5.03	4.74	5.05	5.14	5.59	6.00	5.00	5.67
2D CNN + TCN3	5.07	5.39	4.68	3.72	4.98	4.64	4.81	5.03	5.40	5.72	4.78	5.57
2D CNN + TCN4	4.95	5.25	4.58	3.81	4.71	4.49	4.76	4.83	5.27	5.61	4.76	5.27

Table 8: Online age estimation *tMAE* and *tSTD* for our 2D CNN image-based age estimation model (Bešenić et al., 2022) and the proposed video-based age estimation models (2D CNN + TCN) on the CCMiniVID-A benchmark data according to the online video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
2D CNN	6.10±3.12	6.68±3.58	5.39±2.54	4.43±2.54	6.03±3.33	5.53±2.84	5.78±2.88	5.84±2.84	6.58±3.44	6.98±3.57	5.85±3.02	6.52±3.29
2D CNN + TCN3	5.36±1.79	5.75±2.07	4.88±1.44	4.12±1.70	5.33±1.82	4.91±1.64	5.07±1.68	5.23±1.64	5.73±1.98	6.05±2.03	5.06±1.71	5.85±1.93
2D CNN + TCN4	5.16±1.51	5.52±1.74	4.71±1.23	4.02±1.40	4.97±1.53	4.69±1.40	4.94±1.42	5.00±1.37	5.52±1.67	5.85±1.70	4.96±1.44	5.49±1.62

Our method follows the basic design principle of many video-processing methods (De Geest et al., 2016; De Geest and Tuytelaars, 2018; Xu et al., 2019; Kim et al., 2021; Wang et al., 2022), including some previously reviewed video age estimation methods (Pei et al., 2019; Ji et al., 2018), meaning that the model consists of an image-based 2D CNN feature extractor and a temporal model that learns to aggregate frame-level features in an optimal way.

For feature extraction, we once again rely on our 2D CNN model for image-based age estimation from Section 4.1.1. The feature extractor backbone is stripped of its last age classification layer to produce features of 512 elements. Motivated by the results of (Bai et al., 2018), we base our temporal model on their implementation of Temporal Convolutional Network (TCN)[3]. The proposed TCN can be parameterized with respect to number of layers, kernel size, and number of input and output channels. We set the kernel size to 2 and parameterize the TCN to input and output feature vectors of size 512, matching the chosen feature extractor output. The TCN's output is passed to a fully connected layer that translates the temporal feature vectors to age estimations. The number of layers determines the network's receptive field and therefore the time window size used in online processing. We experiment with 3 and 4 layers, resulting in receptive fields of 8 and 16 frames, respectively.

An overview of the method is presented in Figure 4. The figure illustrates the method's processing pipeline for a 4-layer TCN with a receptive field of 16 frames. The method predicts a softmax probability vector y'_t of size 1 × 101 (mapping to ages from 0 to 100) based on the current frame face crop x_t and 15 previous face crops (x_{t-1} to x_{t-15}). The weighted sum of the softmax probabilities is used as the final age prediction, according to the DLDLv2 method (Gao et al., 2018).

We train the proposed model in an end-to-end manner, meaning that we jointly optimize both the image-pretrained feature extractor and the randomly initialized TCN. The training data is divided into training and validation subsets with a random 80:20 split. The model is trained with the DLDLv2 age estimation method and Adam optimization (Kingma and Ba, 2014), using a learning rate of 10^{-6} and weight decay of 10^{-3}. We adopt an early stopping approach, where training is ended when the validation subset *MAE* plateaus.

5.3.3 Video-Based Benchmark Results

The results in Tables 7 and 8 show that we outperformed image-based baseline according to both offline and online evaluation protocols, even though our model trainings were supervised with pseudo labels generated by that exact baseline model. Figure 3 presents consistent improvements of the 4-layer TCN model (TCN4) on the offline protocol in relation to sequence duration. Significant improvements can also be observed in the online protocol results in Table 8, where the TCN4 overall *tMAE* is reduced by 15.41%, while *tSTD* is reduced by an even larger margin of 51.60%.

The TCN3 model uses a time window of size 8, which amounts to just 1.6 seconds of video data un-

[3] github.com/locuslab/TCN/blob/master/TCN/tcn.py

der the proposed frame subsampling setting. By using the first 1.6 seconds of each video, TCN3 can achieve *MAE* of 5.32 with a single online inference pass, compared to 5.52 in the case of the aggregated image-based results. Image-based result aggregation on the full 20-second videos gives *MAE* of 5.25. By using a single online TCN4 inference pass on the first 3.2 seconds of each video, we even outperform the full-video results with *MAE* of 5.15, simultaneously achieving 84% shorter estimation time and improved estimation accuracy.

6 CONCLUSIONS

Aligned with our initial premise for this work, *taking a better look* on video sequences significantly improves age estimation, compared to the single-image approach. Our evaluation also confirmed previous findings regarding inconsistent performance with respect to gender, skin tone, and lighting type, highlighting the issue of demographic biases in ML. Contrary to the findings of previous work, our experiments showed that the correlation between age estimation error of contemporary age estimation models and auxiliary face tracking data (*i.e.*, head-pose angles, tracking quality, and face scale) is negligible, making the frame cherry-picking method ineffective. The proposed video age estimation method, based on pseudo-labeling and semi-supervised learning, overcomes the lack of available annotated training data and improves age estimation accuracy according to both offline and online evaluation protocols, while estimation stability is improved by more than 50%. Our goal is that our carefully designed and publicly available benchmark protocol, along with the baseline video age estimation results, encourages and supports further research on the topic of video-based age estimation, as the potential of this underexplored field of research is clearly demonstrated.

ACKNOWLEDGEMENTS

This research is funded by Visage Technologies AB.

REFERENCES

Agbo-Ajala, O. and Viriri, S. (2021). Deep learning approach for facial age classification: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54:179–213.

Al-Shannaq, A. S. and Elrefaei, L. A. (2019). Comprehensive analysis of the literature for age estimation from facial images. *IEEE Access*, 7:93229–93249.

Alnajar, F., Shan, C., Gevers, T., and Geusebroek, J.-M. (2012). Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions. *Image and Vision Computing*, 30(12):946–953.

Angulu, R., Tapamo, J. R., and Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):1–35.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bassili, J. N. (1979). Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11):2049.

Bešenić, K., Ahlberg, J., and Pandžić, I. S. (2022). Picking out the bad apples: unsupervised biometric data filtering for refined age estimation. *The Visual Computer*, pages 1–19.

Chen, B.-C., Chen, C.-S., and Hsu, W. H. (2015). Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815.

Chen, J., Mittal, G., Yu, Y., Kong, Y., and Chen, M. (2022). Gatehub: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19934.

Chen, S., Liu, Y., Gao, X., and Han, Z. (2018). Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer.

De Geest, R., Gavves, E., Ghodrati, A., Li, Z., Snoek, C., and Tuytelaars, T. (2016). Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 269–284. Springer.

De Geest, R. and Tuytelaars, T. (2018). Modeling temporal structure with lstm for online action detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1549–1557. IEEE.

Dibeklioğlu, H., Alnajar, F., Salah, A. A., and Gevers, T. (2015). Combining facial dynamics with appearance for age estimation. *IEEE Transactions on Image Processing*, 24(6):1928–1943.

Dibeklioğlu, H., Gevers, T., Salah, A. A., and Valenti, R. (2012a). A smile can reveal your age: Enabling facial dynamics in age estimation. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 209–218.

Dibeklioğlu, H., Salah, A. A., and Gevers, T. (2012b). Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 525–538. Springer.

Eun, H., Moon, J., Park, J., Jung, C., and Kim, C. (2021). Temporal filtering networks for online action detection. *Pattern Recognition*, 111:107695.

Fitzpatrick, T. B. (1975). Soleil et peau. *J. Med. Esthet.*, 2:33–34.

Gao, B.-B., Zhou, H.-Y., Wu, J., and Geng, X. (2018). Age estimation using expectation of label distribution learning. In *IJCAI*, pages 712–718.

Gao, J., Yang, Z., and Nevatia, R. (2017). Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*.

Guo, G., Mu, G., Fu, Y., and Huang, T. S. (2009). Human age estimation using bio-inspired features. In *2009 IEEE conference on computer vision and pattern recognition*, pages 112–119. IEEE.

Hadid, A. (2011). Analyzing facial behavioral features from videos. In *Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2*, pages 52–61. Springer.

Han, H., Otto, C., and Jain, A. K. (2013). Age estimation from face images: Human vs. machine performance. In *2013 international conference on biometrics (ICB)*, pages 1–8. IEEE.

Han, J., Wang, W., Karaoglu, S., Zeng, W., and Gevers, T. (2021). Pose invariant age estimation of face images in the wild. *Computer Vision and Image Understanding*, 202:103123.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., and Ferrer, C. C. (2021). Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332.

Hill, H. and Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current biology*, 11(11):880–885.

Hu, Z., Yang, Z., Hu, X., and Nevatia, R. (2021). Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108.

Ji, Z., Lang, C., Li, K., and Xing, J. (2018). Deep age estimation model stabilization from images to videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1420–1425. IEEE.

Kim, Y. H., Nam, S., and Kim, S. J. (2021). Temporally smooth online action detection using cycle-consistent future anticipation. *Pattern Recognition*, 116:107954.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knight, B. and Johnston, A. (1997). The role of movement in face recognition. *Visual cognition*, 4(3):265–273.

Lee, J.-H., Chan, Y.-M., Chen, T.-Y., and Chen, C.-S. (2018). Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 162–165. IEEE.

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42.

Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., Jain, A. K., Niggel, W. T., Anderson, J., Cheney, J., et al. (2018). Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE.

Monk, Ellis P., J. (2014). Skin Tone Stratification among Black Americans, 2001–2003. *Social Forces*, 92(4):1313–1337.

Niu, Z., Zhou, M., Wang, L., Gao, X., and Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Othmani, A., Taleb, A. R., Abdelkawy, H., and Hadid, A. (2020). Age estimation from faces using deep learning: A comparative analysis. *Computer Vision and Image Understanding*, 196:102961.

O'Toole, A. J., Roark, D. A., and Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in cognitive sciences*, 6(6):261–266.

Panis, G., Lanitis, A., Tsapatsoulis, N., and Cootes, T. F. (2016). Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46.

Pei, W., Dibeklioğlu, H., Baltrušaitis, T., and Tax, D. M. (2019). Attended end-to-end architecture for age estimation from facial expression videos. *IEEE Transactions on Image Processing*, 29:1972–1984.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368.

Porgali, B., Albiero, V., Ryda, J., Ferrer, C. C., and Hazirbas, C. (2023). The casual conversations v2 dataset.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10–17.

Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE.

Ruiz, N., Chong, E., and Rehg, J. M. (2018). Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., and Yuille, A. L. (2018). Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2304–2313.

Wang, W., Peng, X., Qiao, Y., and Cheng, J. (2022). An empirical study on temporal modeling for online action detection. *Complex & Intelligent Systems*, 8(2):1803–1817.

Xu, M., Gao, M., Chen, Y.-T., Davis, L. S., and Crandall, D. J. (2019). Temporal recurrent networks for online action detection. In *IEEE International Conference on Computer Vision (ICCV)*.

Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., and Soatto, S. (2021). Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099.

Zhang, B. and Bao, Y. (2022). Age estimation of faces in videos using head pose estimation and convolutional neural networks. *Sensors*, 22(11):4171.

Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928.

Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., and Loy, C. C. (2022). CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*.

Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155.

APPENDIX A: CCMiniIMG Metadata Reproduction

As discussed in Sections 3.1 and 4.1, the authors of (Hazirbas et al., 2021) extracted 100 frames from each of the 6,022 videos and detected faces with the DLIB face detector (King, 2009). The authors provide only video-level metadata and raw frames. Face detection metadata was not provided, and neither face detection setup nor frame sampling algorithm were specified.

To reproduce the evaluation protocol, we first process CCMiniIMG raw frames following the limited information available in (Hazirbas et al., 2021) and explore DLIB's face detection options. DLIB offers two types of face detectors: HOG-based and CNN-based. The HOG-based face detector is very light, but it is not able to detect many of the challenging samples from this dataset, therefore we choose the CNN-based detector. We experiment with DLIB's sole face detection parameter (i.e., upsampling factor) and CLAHE image enhancement (Pizer et al., 1987) with different tile sizes and clip limits to push the detection rate to 99.82%. No protocol for handling multiple detections was specified, so in each frame, we select the detection with the highest detection confidence. 62 of the 3,011 subjects didn't provide age labels, eliminating 12,400 images from the test set. This results in 588,720 video frames with successfully detected faces and valid age labels. We proceed by extracting face crops with DLIB's affine alignment algorithm which uses 5 face landmarks detected by DLIB's shape predictor. Face crops were extracted with 50% context and resized to 256×256p.

To enable easy reproduction of this evaluation protocol, we make DLIB's detection metadata and CCMiniIMG frame processing scripts publicly available[4] and encourage authors to utilize it in their work to avoid inconsistencies.

APPENDIX B: CCMiniVID Explorative Analysis

All three versions of the CCMiniVID are based on videos from the same 3,011 subjects, with two videos per subject selected from CC to create a balanced subset. Videos of 62 subjects that did not provide age information are not suitable for an age estimation benchmark. That leaves us with 5,898 videos of 2,949 unique subjects used in CCMiniVID benchmark datasets.

The original CC metadata provides some subject-level labels (i.e., age, gender, and skin tone) and video-level labels for lighting (i.e., bright or dark). The authors encourage users to extend the annotations on their dataset (Hazirbas et al., 2021). To this extent, we processed the dataset with a commercial face tracking system[5] and filtered the results by automatically processing the tracking data and manually val-

[4] https://github.com/kbesenic/CCMiniVID

[5] https://visagetechnologies.com/facetrack/

idating edge cases. The tracking data consists of 75 facial landmark points, apparent pitch, yaw, and roll head rotation, and tracking quality for each frame.

CCMiniVID-O label distributions for labels from the original CC metadata are presented in Section 3.2. We can see that gender and lighting label distributions are fairly balanced, with some reasonable underrepresentations in the skin tone distribution. CCMiniVID-A shares the same distribution for the subject-level labels (i.e., age, gender, and skin tone), but there is a minor difference in the lighting distribution caused by the selection of alternative videos (i.e., 62.67% bright and 37.33% dark). Videos of 10 subjects were dropped in the reduced CCMiniVID-R version, causing a very minor variation in all distributions.

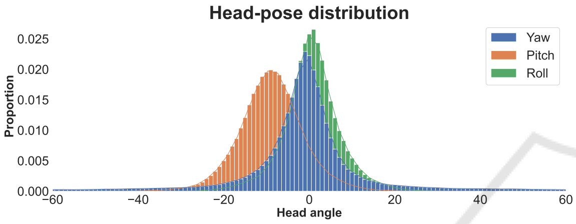

Figure 5: CCMiniVID-O head-pose angle distributions, based on the produced face tracking data.

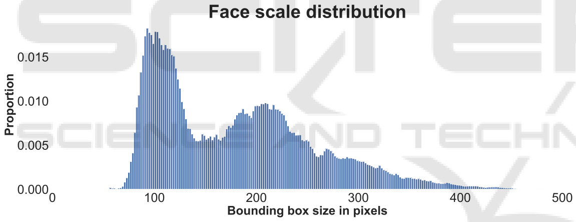

Figure 6: CCMiniVID-O face scale distribution, based on the produced face tracking data.

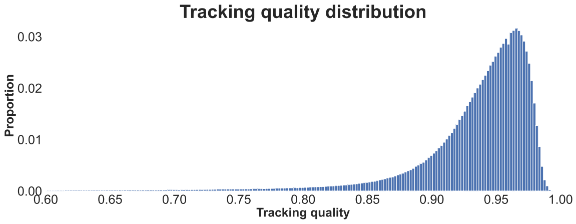

Figure 7: CCMiniVID-O tracking quality distribution, based on the produced face tracking data.

Figures 5, 6 and 7 present distribution for relevant tracking data from our extended CCMiniVID-O metadata. Head-pose angle distributions show that this conversational dataset is oriented towards frontal and near-frontal faces, with some occurrences of extreme head poses. The face scale distribution, where the face scale is the width or height of a square face bounding box in pixels, shows that the provided high-quality videos contain faces that are mostly in the 100 to 400 pixel range. The tracking quality distribution indicates continuous face tracking with high confidence was sustained on a large majority of the videos.