APES: Audiovisual Person Search in Untrimmed Video

Juan León Alcázar[1], Fabian Caba Heilbron [2], Long Mai[2], Federico Perazzi[2],
Joon-Young Lee[2], Pablo Arbeláez[3], and Bernard Ghanem[1]
[1]Universidad de los Andes, [2]Adobe Research, [3]King Abdullah University of Science and Technology,
[1]{juancarlo.alcazar,bernard.ghanem}@kaust.edu.sa [2]{caba,malong,perazzi,jolee}@adobe.com [3]{pa.arbelaez}@uniandes.edu.co

Abstract

Humans are arguably one of the most important subjects in video streams, many real-world applications such as video summarization or video editing workflows often require the automatic search and retrieval of a person of interest. Despite tremendous efforts in the person re-identification and retrieval domains, few works have developed audiovisual search strategies. In this paper, we present the Audiovisual Person Search dataset (APES), a new dataset composed of untrimmed videos whose audio (voices) and visual (faces) streams are densely annotated. APES contains over 1.9K identities labeled along 36 hours of video, making it the largest dataset available for untrimmed audiovisual person search. A key property of APES is that it includes dense temporal annotations that link faces to speech segments of the same identity. To showcase the potential of our new dataset, we propose an audiovisual baseline and benchmark for person retrieval. Our study shows that modeling audiovisual cues benefits the recognition of people's identities.

1. Introduction

Can we find every moment when our favorite actor appears or talks in a movie? Humans can do such search relying on a high-level understanding of the actor's facial appearance while also analyzing their voice [3]. The computer vision community has embraced this problem primarily from a visual perspective by advancing face identification [26, 24, 28]. However, the ability to search for people using audiovisual patterns remains limited. In this work, we address the lack of large-scale *audiovisual* datasets to benchmark the video person retrieval task. Beyond finding actors, several video domain applications could benefit from our dataset, from accelerating the creation of highlight moments to summarizing arbitrary video data via speaker diarization.

Contrary to image collections, video data casts additional challenges for face and person retrieval tasks [11]. Such challenges include drastic changes in appearance, facial expressions, pose, or illumination as a video progresses. These challenges have fostered research in video person search. Some works have focused on person re-identification in surveillance videos [10, 9, 32]. In this setup, the goal is to track a person among a set of videos recorded from various cameras, where the global appearance (*e.g.* clothing) of the target person remains constant. Another setup is the cast search problem, where models take a portrait image as a query to retrieve all person tracks that match the query's identity [14]. The community has achieved relevant progress, but the lack of large scale audiovisual information still prevents the development of richer multi-modal search models.

Motivated by PIN cognitive models [3], Nagrani *et al*. [19] have developed self-supervised models that learn joint face-voice embeddings. Their key idea is to use supervision across modalities to learn representations where individual faces match their corresponding voices via contrastive learning [12]. However, many videos in the wild might contain multiple visible individuals that remain, mostly, silent. This situation introduces a significant amount of noise to the supervision signal [25]. Liu *et al*. [17] have also explored audiovisual information for person retrieval in videos. To this end, their work introduces the iQIYI-VID dataset, which contains videos from social media platforms depicting, in large proportion, Asian celebrities. Despite its large-scale, the dataset contains only short clips, with most of them being five seconds long or shorter. We argue that having long videos is crucial to high-level reasoning of context to model real-life expressions of people's faces and voices. Additionally, we require densely annotated ground-truth labels to enable direct links between speech and visual identities.

In this paper, we introduce APES (Audiovisual PErson Search), a novel dataset, and benchmark for audiovisual person retrieval in long untrimmed videos. Our work aims to mitigate existing limitations from two angles. First, we densely annotate untrimmed videos with person identities and match those identities to faces and voices. Second, we establish audiovisual baselines and benchmarks to facilitate

Figure 1. **Audiovisual PErson Search (APES).** We introduce APES, a novel video dataset for multimodal person retrieval. The dataset includes annotations of identities for faces and voice segments extracted from untrimmed videos. We establish three new person search tasks. (i) Seen aims to retrieve all timestamps when a target person is on-screen. (ii) Seen & Heard, which focuses on finding all instances when a person is visible and talks.

future research with the new dataset. Our dataset includes a broad set of 15-minutes videos from movies labeled among a long-tailed distribution of identities. The dataset samples account for many challenging re-identification scenarios such as small faces, poor illumination, or short speech segments. In terms of baselines, we develop a two-stream model that predicts people's identities using audiovisual cues. We include benchmarks for two alternative tasks. *Seen*, which aims at retrieving all segments when a query face appears on-screen; *Seen & Heard*, which focuses on finding instances where the target person is on-screen and talking. Figure 1 showcases APES annotations and tasks.

Contributions. This paper's primary goal is to push the envelope of audiovisual person retrieval by introducing a new video dataset annotated with face and voice identities. To this end, our work brings two contributions.[1]

- We annotate a dataset of 144 untrimmed videos from movies. We associate more than $30K$ face tracks with $26K$ voice segments and label about $1.9K$ identities. Section 3 details our data collection procedure and showcases the characteristics of our novel dataset.
- We establish an audiovisual baseline for person search in Section 4. Our study showcases the benefits of modeling audiovisual information jointly for video person retrieval in the wild.

2. Related Work

There is a large corpus of related work on face [28, 26, 24] and voice [30, 4] retrieval. This section focuses on related work on datasets for video person retrieval and audiovisual learning for identity retrieval.

[1]The dataset, and the code to reproduce baselines and benchmarks will be released soon.

Video Person Retrieval Datasets. After many milestones achieved on image-based face retrieval, the computer vision community shifted attention into video use cases. There are many datasets and tasks related to person and face retrieval in videos. Three popular tasks have been established, including person re-identification, speaker recognition, and recently person search. Table 1 summarizes datasets for these tasks and compares them with the APES dataset.

The first group includes datasets designed for person re-identification [32, 31, 29]. These datasets usually contain many identities; however, most are composed only of cropped tracks without any visual context or audio information. Moreover, person re-identification datasets focus on surveillance scenarios, where the task is to find a target subject across an array of cameras placed in a limited area.

The second group includes speaker recognition datasets [4, 21]. Datasets such as VoxCeleb2 [4] have pushed the scale up to 150K face clips. A drawback of this group of datasets is that clips are only a few seconds long and tend to contain a single face. The third group includes datasets for person retrieval. CSM [14], for instance, introduces the task of cast search, which consists of retrieving all the tracklets that match the identity of a portrait query. iQIYI-Video [17] scales the total number of tracklets, clips, and identities. Both datasets provide a step towards visual-based person retrieval but exhibit limitations for multimodal (faces and voices) modeling. On the one hand, CSM does not provide audio streams or video context; on the other hand, iQIYI-Video contains short clips and does not associate voices with person identities. Our APES dataset mitigates these limitations by annotating long videos with people's faces, voices, and their corresponding identities.

The third group is the closest to our setup, it comprises datasets for audiovisual person search. While the Big Bang Theory datatset [2] allows to study the same tasks as APES,

Dataset	Source	Task	# Instances	# Tracklets	# Identities
PSD [29]	Images	Re-identification	96K	-	8432
Market [32]	Images	Re-identification	32K	-	1501
MARS [31]	Person Tracks	Re-identification	1M	20K	1261
VoxCeleb [21]	Short Clips	Speaker Recognition	-	22.5K	1251
VoxCeleb2 [4]	Short Clips	Speaker Recognition	-	150K	6112
iQIYI-VID [17]	Short Clips	Visual Search	70M	600K	5000
CSM [14]	Person Tracks	Visual Search	11M	127K	1218
Big Bang Theory [2]	Untrimmed Videos	Audiovisual Search	-	3.7K	8
Buffy [8]	Untrimmed Videos	Audiovisual Search	49k	1k	12
Sherlock [22]	Untrimmed Videos	Audiovisual Search	-	6.5K	33
APES	Untrimmed Videos	Audiovisual Search	3.1M	30.8K	1913

Table 1. **Video person identity retrieval datasets overview.** APES is the largest dataset for audiovisual person search. In comparison to available audiovisual search datasets, it contains two orders of magnitude more identities at 1.9K. Additionally, the 30K manually curated face tracks contain a much larger diversity in audiovisual patterns than similar datasets, as its original movie set is composed of far more diverse videos across multiple genres and including diverse demographics. Finally, the 3.1 Million individual instances allow for modern machine learning methods techniques to be used on the APES dataset.

it is limited to only 8 identities (which are observed mostly indoors). Additionally speech events are approximately localized using the show's transcripts. APES contains dense manual annotations for speech events and identity pairs. Sherlock [22] also allows for audiovisual identity search but contains only 33 identities, and its cast is composed of mostly white European adults. The Sherlock dataset also discards short segments of speech (shorter than 2 secs), this is a key limitation as our analysis shows that these short segments constitute a big portion of utterances in natural conversations. Finally, Buffy [8] is also very small in terms of number of identities and its data lacks diversity as it was collected from only two episodes of the series.

Audiovisual Learning for Identity Retrieval. Audiovisual learning has been widely explored in the realm of multiple video tasks [23, 5, 25, 1], but only a few have focused their efforts on learning embeddings for person identity retrieval [20, 19]. Nagrani *et al*. [20] have proposed a cross-modal strategy that can 'see voices' and 'hear faces'. It does so by learning to match a voice to one out of multiple candidate faces using a multi-way classification loss. More recently, the work in [19] introduces a cross-modal approach that leverages synchronization between faces and speech. This approach assumes there is one-to-one correspondence in the audiovisual signals to form queries, positive, and negative sets and train a model via contrastive learning [12]. Although the method proposed in [19] does not require manually annotated data, it assumes all face crops contain a person talking, an assumption that often breaks for videos in the wild. Our baseline model seizes inspiration from the success of these previous approaches in cross-modal and audiovisual learning. It leverages the newly annotated APES dataset, and its design includes a two-stream audiovisual network that jointly learns audiovisual characteristics of individuals.

3. APES: Audiovisual PErson Search Dataset

This section introduces the Audiovisual PErson Search (APES) dataset, which aims at fostering the study of multimodal person retrieval in videos. This new dataset consists of more than 36 hours of untrimmed videos annotated with 1913 unique identities. APES' videos pose many challenges, including small faces, unconventional viewpoints, as well as short segments of speech intermixed with environmental sound and soundtracks. Figure 2 shows a few APES samples. We plan to release the dataset with its full annotation set and the baseline models to foster the research of audiovisual models for person retrieval. Here, we describe our data collection procedure and statistics of APES.

3.1. Data Collection

Video source. We aim for a collection of videos showing faces and voices in unconstrained scenarios. While there has been a surge of video datasets in the last few years [15, 18, 27, 6], most of them focus on action recognition on trimmed videos. As a result, it is hard to find a large video corpus with multiple instances where individuals are seen on-screen and speaking. This trend limits the availability of relevant data for audiovisual person search.

Instead of gathering user-generated videos, the AVA dataset [11] is made from movies. AVA list of movies includes productions filmed around the world in different languages and across multiple genres. It contains a diverse and extensive number of dialogue and action scenes. Another

Figure 2. **The APES dataset.** We illustrate a few examples from our novel APES dataset. As it can be seen, it casts many challenges for automatic audiovisual person identification and search. For instance, the appearance of each identity changes significantly across different scenes in a movie. Similarly, APES poses challenges to identify voices across time, as the environment and background sounds are always changing. In many cases, even detecting the persons' faces can be challenging due to illumination and partial occlusions.

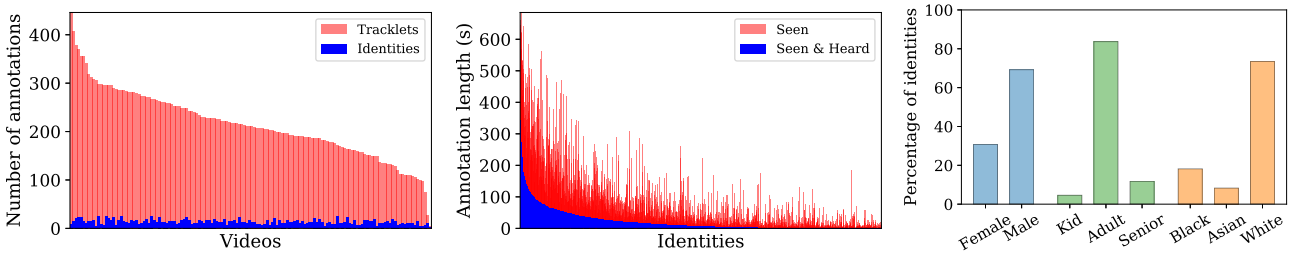

Figure 3. **APES global statistics.** *Left:* Distribution of number of tracklets and identities per video, the length of the tracklets roughly follows a long tail distribution, but there is not much difference in the number of identities per video. *Center:* Distribution of annotation length of only seen identities, and seen & heard identities, it is important to note that many actors are seen but remain silent, and the speaking characters get most screen time. *Right:* Demographics of labeled identities in APES, sorter by gender, age group, and race. Despite representing demographics better than previous datasets [2, 17], we still observe room for improvement towards a more balanced distribution. Nevertheless, for most demographics, the dataset contains at least 1.5K tracks and 155K face crops, a representative set for training deep learning models.

appealing property is that it provides a download mechanism [2] to gather the videos in the dataset; this is crucial for reproducibility and promote future research. Finally, the AVA dataset has been augmented with Active Speaker annotations [25]. This new set contains face tracks annotated at 20 frames per second; it also includes annotations that link speech activity with those face tracks. Consequently, we choose videos from AVA to construct the APES dataset. Our task is then to label the available face tracks and speech segments with actors' identities.

Labeling face identities. We first downloaded a total of 144 videos from the AVA dataset, gathered all face tracks available, and *annotated* identities for a total of 33739 face tracks. We did so in two stages. In the first stage, we addressed the identity assignment tasks per video and asked human annotators to cluster the face tracks into matching identities. To complete this first task, we employed three human annotators for 40 hours each. We noticed two common errors during this stage: (i) false positives emerging from small faces or noisy face tracks; and (ii) false negatives that assign the same person into more than one identity cluster. We alleviated these errors by implementing a second stage to review all instances in the clustered identities and merged wrongfully split clusters. This process was a relatively shorter verification task, which annotators completed in *eight* hours. At the end of this annotation stage, about 8.4% of the face tracks were labeled as ambiguous; therefore, we obtained a total of 30880 face tracks annotated among 1913 identities.

Labeling voice identities. After labeling all face tracks with their corresponding identities, we now need to find their voice pairs. Our goal then is to cluster all voices in the videos and match their corresponding faces' original identity. In other words, we want the same person's faces and voices to share the same identity. Towards this goal, we leveraged the original annotations from the AVA-ActiveSpeakers dataset [25], which contain temporal win-

[2] Researchers can download the full set of AVA videos at: https://github.com/cvdfoundation/ava-dataset

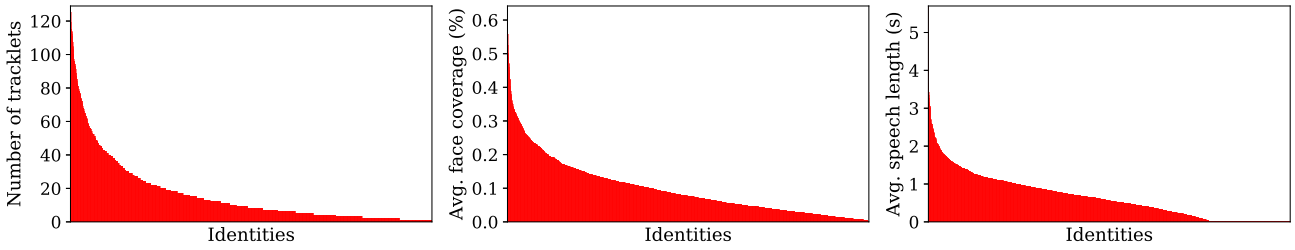

Figure 4. **APES identity statistics.** We analysis 3 relevant statistics for the label distribution in APES, all of them follow a long tailed distributions *Left:* Number of face tracks per identity. This distribution indicates that main characters tend to appear more often than, for instance, extras or background actors. *Center:* Average face coverage per identity, computed as the mean of the relative area covered by all identity face tracks. This statistic hints that a big portion of identites are framed within close-ups, but still many identities are portrait in the background of the scene. *Right:* Average length of continuous speech per identity. Interestingly, most speech segments last at most four seconds, and about 25% of the identities do not speak at all.

dows of speech activity associated with a face track. We mapped speech segments to their corresponding face track and assigned a common identity. We annotated 26879 voice segments accounting for a total of 11.1 hours of speech among the 1913 identities.

3.2. Dataset Statistics

We annotated over 36 hours from 144 videos, including 30880 face tracks, $3.1M$ face bounding boxes, and 26879 voice segments. Our labeling framework annotated 1913 identities and discarded 2859 ambiguous face tracks. We discuss in detail the statistics of the dataset below.

Global statistics. In Figure 3 (*Left*), we observe the distribution of the number of tracklets and identities per video. The number of tracklets per video follows a long-tail distribution, and there is no correlation between the number of tracklets with the number of identities. This fact indicates that certain identities have longer coverage than others. Figure 3 (*Center*) shows the average length an identity is Seen or Seen & Heard. Interestingly, the Seen & Heard distribution exhibits a long-tailed distribution, with many identities being heard only very few times. Also, some identities are seen many times without speaking at all. Finally, we investigate the demographics of the dataset in Figure 3 (*Right*). To do this, we manually annotate the identities with gender, age, and race attributes. Although work needs to be done to balance samples across demographics, the survey shows that our video source has representative samples to cover the various demographic groups. For instance, for the most under-represented group, kids, APES contains more than 1.5K tracks. Moreover, APES provides significant progress from previous datasets that contain a single demographic group, *e.g.*, iQIYI-VID [17] contains only Asian celebrities, and the Big Bang Theory dataset [2] comprises a cast limited to a single TV series.

Identity statistics. We analyze here characteristics of the annotated identities. First, in Figure 4 (*Left*), we show the distribution of the number of tracklets per identity. We observe a long-tailed distribution where some identities, likely the main characters, have ample screen time, while others, *e.g.* supporting cast, appear just a few times. Figure 4(*Center*) shows the average face coverage per identity, where we observe also a long-tail distribution. On the one hand, identities with large average face coverage include actors favored with close-ups; on the other hand, identities with low average face coverage include actors framed within a wide shot. Finally, we plot the average length of continuous speech per identity (Figure 4 (*Right*)). Naturally, different characters have different speech rhythms, and therefore the dataset exhibits a non-uniform distribution. Interestingly a big mass centers around one second of speech. This characteristic might be due to the natural dynamics of engaging dialogues. Moreover, we observe than about 25 % of the identities do not speak at all.

4. Experimental Analysis

4.1. Baseline methods.

We now outline the standard evaluation setup and metrics for the APES dataset along with a baseline method that relies on a two-stream architecture for multi-modal metric learning. The first stream receives cropped faces while the second works over audio clips. Initially, each stream is optimized via triplet loss to minimize the distance between matching identities in the feature space. As highlighted in other works [26, 20, 24], it is essential to acquire a clean and extensive set of triplets to achieve good performance. Below, we detail each modality of our baseline model and different subsets for training.

Facial Matching. To optimize the face matching network, we remove the classification layer from a standard

Sampling strategy	Seen & Heard Avg Positive Tracklets	Seen & Heard Avg Negative Tracklets	Seen & Heard Clean	Seen Avg Positive Tracklets	Seen Avg Negative Tracklets	Seen Clean
Weak	1	11862	✗	1	23679	✗
Within	10	112	✓	16	242	✓
Across	10	11862	✓	16	23679	✓

Table 2. **Identity Sampling.** APES allows for multiple sampling strategies to form triplets during training. APES-IN is the simplest scenario as it samples positive and negative identities from a single video. Such sampling generates a 10:112 positive-to-negative ratio, for instance, in the case of the Seen & Heard task. APES-Across creates a more challenging scenario, where negatives are extracted from all videos across the dataset; it is much more imbalanced problem with a 10:11862 positive-to-negative ratio. APES-weak considers an extreme scenario where a single tracklet is used to retrieve identities from the whole video collection; this setting is not only extremely unbalance 1:11862 but also challenging as the audiovisual identity from a single tracklet tend to be highly similar in appearance and sound characteristics.

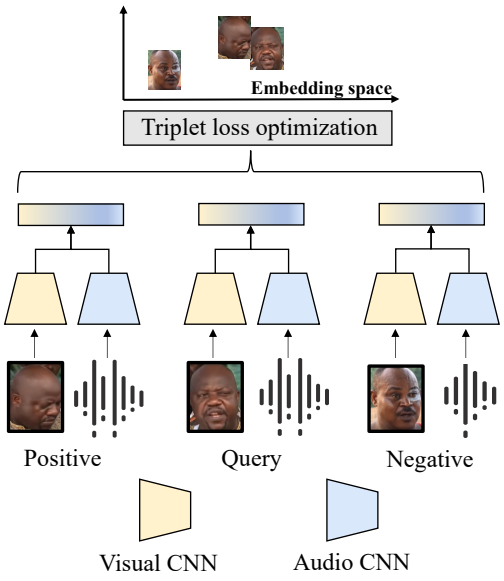

Figure 5. **Creating Audiovisual Embeddings** , We use a two stream neural network to identify corresponding identities in APES. Our approach uses independent visual (light yellow) and audio (light blue) CNNs, the joint feature set is then optimized by means of the triplet loss to obtain an embedding where corresponding identities are close located,

Resnet-18 encoder [13] pre-trained on ImageNet [7], and fine-tune it using a triplet loss [26]. We choose the ADAM optimizer [16] with an initial learning rate of 3×10^{-4} and learning rate annealing of 0.1 every 30 epochs for a total of 70 epochs. We resize face crops to 124×124 and perform random flipping and corner cropping during training.

Voice Matching. Similar to the visual stream, we use a ResNet-18 model initialized with ImageNet weights and fine-tuned via triple loss learning. We follow a setup similar to [25] and use a Mel-spectrogram calculated from audio snippets of 0.45 seconds length in the audio stream. We use the same hyper-parameters configuration described for the visual matching network.

Cross-modal Matching. For the audiovisual experiments, we combine the individual configurations for voice and face matching. However, we add a third loss term which optimizes the feature representation obtained from a joint embedding of audiovisual features, which we obtain via concatenation of each stream's last layer feature map. This third loss is also optimized using the triplet loss. Figure 5 illustrates our cross-modal baseline.

4.2. Experimental Setup

Dataset splits and Tasks. We follow the official train-val splits defined in the original AVA-ActiveSpeaker dataset [25]. As not every person is actively speaking at every moment, we define two tasks: *Seen & Heard* (a person is on-screen and talking), and only *Seen* (the person is on-screen but not speaking). Each of these tasks yield a corresponding training and validation subset. The Seen task subsets have 23679 and 7989 tracklets for training and validation respectively. Conversely, the Seen & Heard subset, is comprised of 11862 tracklers for training and 3582 for validation.

Identity Sampling. The APES dataset allows us to sample positive and negative samples during training in three different ways. The most direct sampling would gather every tracklet from a single movie. In such a scenario, we will create a positive bag with the tracklets that belong to a given identity, while negative samples would be obtained from every other tracklet in the same movie. We name this setup **Within**, a simplified configuration where we have a 1:15 ratio of positive tracklets (same identity) to negative tracklets (different identity).

While the **Within** modality allows us to explore the problem, it might be an overly simplified scenario. Hence, we also devise the **Across** setup, where we sample negatives identities over the full video collection, *i.e.*, across different movies. This sampling strategy significantly changes the

ratio of positive to negative tracklets to 1:150 and better resembles the natural imbalance of positive/negative identities in real-world data.

Finally, we create an extreme setup that resembles few-shot learning scenarios for identity retrieval learning. In this case, a single (or very few) positive samples are available to train our embeddings. These positive samples are sampled from the same tracklet as the query, and instances from every other tracklet in the datasets form the negative bag. We name this setup as **Weak**, which results in a strongly imbalanced subset with about 1:1500 ratio of positive to negative tracklets. A summary of these three sampling sets is presented in Table 2.

Evaluation metrics. Three evaluation metrics assess methods' effectiveness in APES:

- Precision at K (P@K): we estimate the precision from the top K retrieved identities for every tracklet in a video. As there are no shared identities over videos, we simply estimate the precision at K for every video, and then average for the full validation set.
- Recall at K (R@K): we estimate the recall from the top K retrieved identities for every tracklet in a video. Again, we estimate the recall at K for every video and compute the average over the full validation set.
- Mean average precision (mAP): as final and main evaluation metric, we use the mean average precision. Like in the recall and precision cases, we compute the mAP for every tracklet in a video, and the average the results for videos in the validation set.

4.3. Benchmark Results

Seen & Heard benchmark results. Table 3 summarizes our benchmark results for the three main configurations: Facial Matching, where we operate exclusively on visual data; Voice Matching, where we exclusively model audio data; and Cross-modal Matching, where we train and validate over the audio and visual modalities. Overall results obtained with facial and cross-modal matching are far from perfect. Our best baseline model obtains a max mAP of 64.8%. This result indicates that the standard triplet loss is just an initial baseline for the APES dataset and that there is ample room for improvement. The relatively high precisions at K=1 and K=5 in almost every setting, suggests the existence of a few easy positives matches for every query in the dataset. However, the relative low precision and recall at K=10 suggest that the method quickly exhibits wrong estimations as we progress through the ranking. This drawback is worst in the APES-Weak training setting, as its selection bias induces a much lower variability on the positive samples. The analysis of the recall scores highlights the importance of the additional audio information. After the network is enhanced with audio data, the recall metrics improve significantly, reaching 98.5% at K=100. While this improvement comes at the cost of some precision, overall, the mAP shows an improvement of 0.7%. Finally, there is only a slight difference between the Within and Across sampling settings, the former having a marginally better recall. This suggests that the massive imbalance induced in the across setting does not significantly improve the diversity of the data observed at training time and that more sophisticated sampling strategies such as hard negative mining might be required to take advantage of that extra information.

	R@10	R@50	R@100	P@1	P@5	P@10	mAP(%)
Random	12.0	52.9	87.3	18.1	18.7	18.4	19.1
Facial Matching							
Weak	38.2	74.5	94.2	80	62.3	49.9	45.4
Within	48.1	86.5	98.1	87.5	**74.0**	**62.8**	64.1
Across	**48.2**	86.3	98.2	**88.0**	73.6	62.6	63.4
Voice Matching							
Weak	17.4	59.1	89.4	29.2	26.9	23.2	20.5
Within	18.2	59.7	89.9	30.0	26.7	24.8	21.8
Across	17.7	58.9	89.5	29.3	26.6	24.5	22.0
Cross-modal Matching							
Weak	38.2	74.2	94.3	80.1	62.0	49.4	47.9
Within	47.9	**87.1**	**98.5**	85.9	72.4	61.2	**64.8**
Across	48.1	86.8	98.3	86.7	73.3	62.2	63.6

Table 3. **Benchmark results for Seen & Heard task.** We measure recall at K (R@K), precision at K (P@K), and the mean Averge Precision (mAP). We observe that the Cross-modal matching baseline slightly outperforms the facial matching model (by 0.7% mAP). This result suggests that the audio signal is helpful to the re-identification process, especially increasing the recall of the proposed method. The relative low precision for larger K's suggest that correspondences are easy to build for a few images, but much harder as candidates become more diverse. We also observe that identifying persons using only voice is a much harder task.

	Seen Full Set						
	R@10	R@50	R@100	P@1	P@5	P@10	mAP(%)
Random	5.1	26.0	52.7	17.2	17.1	16.9	17.1
Facial Matching							
Weak	26.9	54.5	74.5	82.2	66.0	54.6	39.3
Within	33.1	70.5	87.4	87.8	77.3	68.7	58.8
Across	33.8	70.4	87.6	88.2	77.6	68.6	59.4

Table 4. **Benchmark for the Seen Set.** We use the same metrics as in the Seen & Heard setting. We find that this configuration is slightly harder despite having more data. This counter-intuitive result can be explained as movies often rely on shots where a active-speakers are portray in close-ups. In short, the Seen task has more data available for training, but also more challenging scenarios.

Seen benchmark results. We empirically found that fusing modalities with noisy audio data (original splits) provides no improvement. As outlined before, our experiments suggest that audio models are highly sensitive to

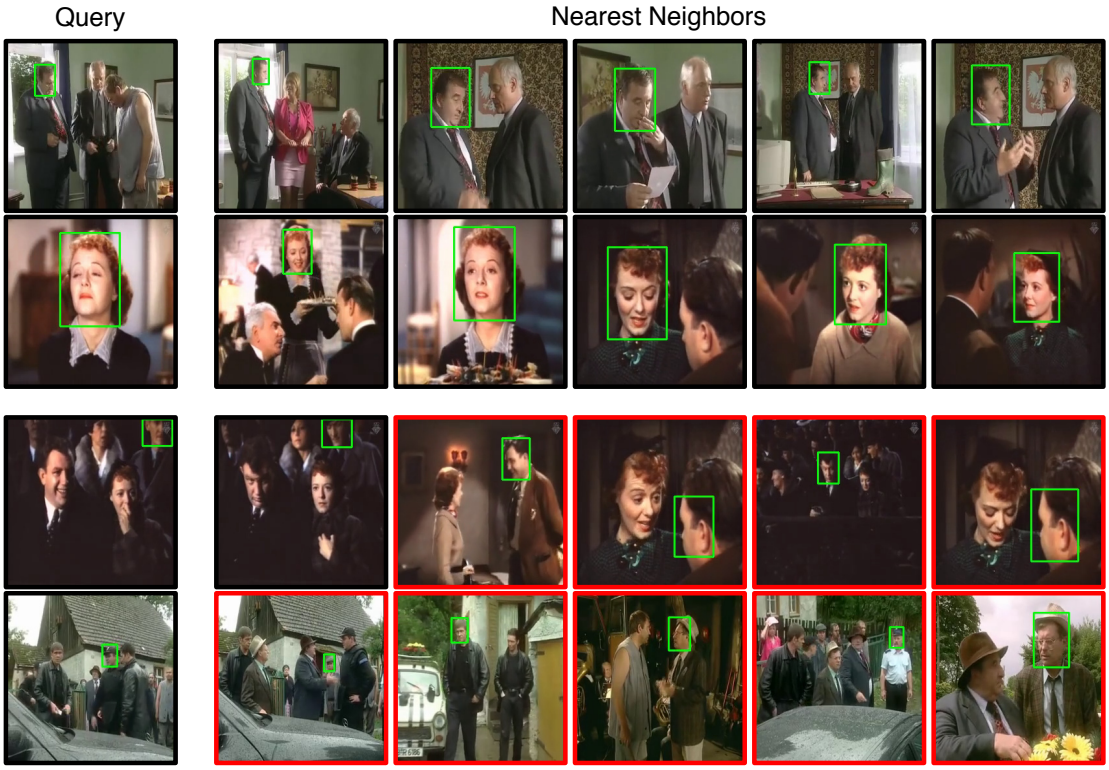

Figure 6. **Qualitative results.** We showcase easy (top two rows) and hard (bottom two rows) examples in our benchmark. Green bounding boxes indicate the localization of the queried and retrieved faces, and red borders around the frame indicates a false positive. We empirically observe that queries from large faces within dialogue scenes tend to be easier to retrieve. In contrast, queries from small faces and cases where the subject is moving are challenging to our baseline.

noisy speech annotations and do not converge if there is large uncertainty in the ground truth. Under this circumstances the optimization just learns to ignore the audio cues, and yields the same performance as the visual-only setting.

Despite this we report our results of a facial matching for the Seen task, as it will serve as baseline for future works that can handle noise speech data. Table 4 contains the benchmark for the Seen task. We find that this task is actually harder than the Seen & Heard task, despite having more available data. We explain this result as movies typically depict speakers over large portions of the screen and offer a clear view angle to him/her, which results on average larger faces with less noise in the Seen & Heard task, and smaller more challenging faces in the Seen. In other words, we hypothesize that speaking faces are easier to re-identify as they are usually framed within close-ups. This bias makes it harder to find every matching tracklet for the identity (thus reducing recall).

4.4. Qualitative Results

We showcase easy and hard instances for our baseline in Figure 6. Every row shows a query in the left, and the top five nearest neighbors on the right. The first two rows, shows instances where our baseline model correctly retrieves instances of the same person. We have empirically noticed that instances where query faces are large, *e.g.* in close-ups and medium shots from dialogue scenes, our baseline model tends to provide a very good ranking to the retrieved instances. The two rows from the bottom illustrate hard cases where the baseline model fails. We have seen that small faces, poor illumination, occlusion, and subjects in motion present a challenging scenario to our baseline.

5. Conclusion

We introduce APES, a new dataset for audiovisual person search. We compare APES with existing datasets for person identity analysis and show that it complements previous datasets in that those have mainly focused on visual analysis only. We include benchmarks for two tasks *Seen* and *Seen & Heard* to showcase the value of curating a new audiovisual person search dataset. We believe in the crucial role of datasets at measuring progress in computer vision; therefore, we are committed to releasing APES to enable further development, research, and benchmarking in the field of audiovisual person identity understanding.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 3

[2] Martin Bauml, Makarand Tapaswi, and Rainer Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609, 2013. 2, 3, 4, 5

[3] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986. 1

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2, 3

[5] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 3

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6, 2006. 3

[9] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010. 1

[10] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535. IEEE, 2006. 1

[11] Patrick J Grother, Mei L Ngan, and George W Quinn. Face in video evaluation (five) face recognition of non-cooperative subjects. Technical report, 2017. 1

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 1, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Qingqiu Huang, Wentao Liu, and Dahua Lin. Person search in videos with one portrait through visual and temporal links. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 425–441, 2018. 1, 2, 3

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Yuanliu Liu, Bo Peng, Peipei Shi, He Yan, Yong Zhou, Bing Han, Yi Zheng, Chao Lin, Jianbin Jiang, Yin Fan, et al. iqiyivid: A large dataset for multi-modal person identification. *arXiv preprint arXiv:1811.07548*, 2018. 1, 2, 3, 4, 5

[18] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018. 3

[19] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–88, 2018. 1, 3

[20] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018. 3, 5

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 2, 3

[22] Arsha Nagrani and Andrew Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. *arXiv preprint arXiv:1801.10442*, 2018. 3

[23] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 1, 2, 5

[25] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1, 3, 4, 6

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2, 5, 6

[27] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 3

[28] Makarand Tapaswi, Marc T Law, and Sanja Fidler. Video face clustering with unknown number of clusters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5027–5036, 2019. 1, 2

[29] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 2, 3

[30] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6301–6305. IEEE, 2019. 2

[31] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 2, 3

[32] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 1, 2, 3