

Automated Video Labelling: Identifying Faces by Corroborative Evidence

Andrew Brown, Ernesto Coto, Andrew Zisserman

Visual Geometry Group, Department of Engineering Science, University of Oxford
Oxford, England

{abrown,ecoto,az}@robots.ox.ac.uk

Abstract—We present a method for automatically labelling all faces in video archives, such as TV broadcasts, by combining multiple evidence sources and multiple modalities (visual and audio). We target the problem of ever-growing online video archives, where an effective, *scalable* indexing solution cannot require a user to provide manual annotation or supervision. To this end, we make three key contributions: (1) We provide a novel, simple, method for determining if a person is famous or not using image-search engines. In turn this enables a face-identity model to be built reliably and robustly, and used for high precision automatic labelling; (2) We show that even for *less-famous people*, image-search engines can then be used for *corroborative evidence* to accurately label faces that are named in the scene or the speech; (3) Finally, we quantitatively demonstrate the benefits of our approach on different video domains and test settings, such as TV shows and news broadcasts. Our method works across three disparate datasets without any explicit domain adaptation, and sets new state-of-the-art results on all the public benchmarks.

Index Terms—video annotation, person identification;

I. INTRODUCTION

There has been an exponential growth in the volume of video content (in the form of TV and film material) being produced and released online. Such content is rich with useful information for researchers, historians and the general public. However, the sheer scale of the data, coupled with a lack of relevant metadata, makes indexing, analysing and navigating this content an increasingly difficult task. Relying on additional, manual human annotation is no longer feasible, and without an effective way to navigate these videos, this bank of knowledge is largely inaccessible.

Interestingly, most video indexing and analysis is *human-centric*. For example, we might want to navigate to a scene where two particular people interact, or where a group of people first appear together. This is partly because many videos in online archives are centred on humans, but also because of our natural interest in human actions and interactions. The focus of this paper is therefore the labelling of all faces in videos, in a way that does not require any additional manual annotation from a user, be it in the form of provided transcripts [4], [23], [24], [38], [48], or a list of appearing people known *a-priori* [37], [40]. Our approach is hence scalable to large video archives where collecting manual annotations is infeasible.

To solve this task, we take a human-inspired approach. Imagine that you are watching a video and encounter a new person. In order to confidently identify them, you would first



Fig. 1. Modern, large, unlabelled online video archives, such as TV broadcasts, are growing at an exponential rate. These important resources are inaccessible due to the lack of annotations. Our proposed method for automated video labelling is scalable to these large archives, as it does not require additional manual supervision to label faces in videos. Examples from the BBC Videos and Sherlock dataset are shown above. The method is able to label a wide range of people, over different domains, lighting conditions and in extreme poses.

look for clues of their name either in the video such as text on the screen, their name being mentioned in speech, or in a list of cast members from an internet archive. You might then find some evidence to verify that this name is correct, by searching for the person online. In this work we follow this approach by harnessing the freely-available weak annotations on the internet, such as IMDB names lists and image-search engine rankings, to provide evidence for recognising faces automatically and with confidence. Consequently, this method is applicable to identities for whom there are images online.

We denote people with many images of themselves online as *famous*, and introduce a novel approach for automatically identifying if an identity is *famous* without any additional manual supervision or annotation. For identities that are *less-famous* according to our approach, we present a novel method for using other identifying clues in the video together with image-search engine results as *corroborating evidence* to provide confident person labels.

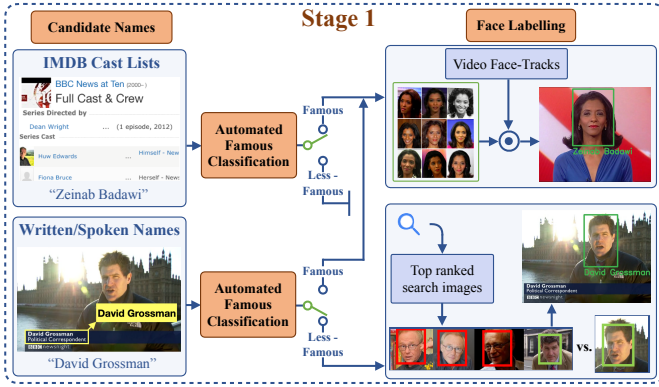


Fig. 2. The first stage of our method automatically finds candidate names and labels faces. Candidate names are automatically sourced from IMDB name lists, displayed text (*written names*) and spoken words. Each person is classified as *famous* or not. For the *famous* people, face-identity models are automatically assembled and they are labelled throughout the videos. For the *less-famous* people who were found in written or spoken names, the temporal occurrence combined with *corroborating evidence* from image-search engines is used to provide labels.

Beyond visual content, videos also contain identity information in the audio track, as humans can be recognised from the sound of their voice. Occasionally when used independently, neither face-appearance or voice can provide a confident identity label, be it due to a slightly obscured face, or a noisy or brief audio signal [37]. By fusing the two modalities and using a separate query expansion step, we show faces can be labelled with confidence, such that the recall of the automated labelling is improved without sacrificing precision. Figure 1 gives example results of our automated labelling method.

In summary, the task of this paper is to assign identity labels (*tags*) to all people in a set of test videos, without the use of any additional manual annotation beyond what is freely and automatically available on the internet. The base unit for labelling used in this paper is the face-track *i.e.* face-detections from consecutive frames of the same identity that are linked together within shots. The method consists of two key stages: **Stage 1 – Using image-search engines as sole/corroborating evidence.** This stage automatically identifies candidate names that may appear in the test videos and classifies them as *famous* or not. Image-search engines are used as the sole source of evidence if they are *famous*, or as *corroborating evidence* if they are not in order to tag the faces. This is presented in Section III and shown in Figure 2.

Stage 2 – Boosting the number of tags. This increases the number and variety of face tags across the test videos, using two techniques: (i) fusing the information from the modalities of face-appearance and voice; and (ii) query expansion. These techniques improve label recall, while maintaining very high precision (see Section IV).

Our method for person labelling can proceed completely automatically starting from just the names of the programmes being labelled (and will obtain lists of appearing names in the process). We can also test specific modules of the method on standard benchmarks where cast lists are provided. We

quantitatively demonstrate the benefits of our approach on several different video domains and test settings, such as TV shows and news broadcasts, as described in Section V. The results in Section VI show that our method works across these disparate datasets without any explicit domain adaptation, and sets new state-of-the-art results on all the public benchmarks. Further details can be found at https://www.robots.ox.ac.uk/~vgg/research/person_id_in_video/.

II. RELATED WORK

Labelling People in Videos: This task has been well studied [4], [23], [24], [27], [37], [38], [40], due to its uses for story understanding [3] and archive indexing. In previous works, different levels of prior information are assumed to be available: Everingham *et al.* [23] make use of transcripts aligned with subtitles to provide weak supervision, with many other works following suit [4], [7], [20], [21], [24], [38], [39], [48], [49]. Often the task is posed as one of Multiple Instance Learning (MIL) [7], [27], [31], [38], [51]. Nagrani and Zisserman [37] instead presume the existence of cleaned web-downloaded images for actor-level supervision. Many of these methods are tasked with labelling only a small number of known, main characters, and either require transcripts or some additional manual annotation in the pipeline. Our method, on the other hand, does not require either, and so crosses the domain gap to real-world large video archive scenarios.

The labelling of people in news videos is also a well studied and challenging task, due to its open-ended nature, where a list of appearing people or transcripts is not available [33], [46], [52]. Both Canseco *et al.* [12] and Maclair *et al.* [30], [36] establish correspondences between spoken names and speech-turns to label people. Several works use the co-occurrence of overlaid names in a scene with speech-turns to make person labels [5], [26], [43], [50], with much of this area of research accelerated by the MediaEval “Person Discovery in Broadcast TV” challenges [44]. However, many of these techniques rely upon heuristics based on TV-broadcast structure to make confident labels. These heuristics do not generalise well to other domains. Our method on the other hand does not rely on such heuristics, and can hence be applied to movies, TV material and broadcast news alike.

Using Image-Search Engines for Supervision: Although widely used in the Vision Community for object recognition [6], [14], [25], [34], [47], and more recently for face recognition [10], [15], [28], [32], [37], [40]–[42], the problem faced is that retrieved results have varying precision. Previous person-identification work has either relied on manually removing false positives from retrieved images [37], or has focused on automatically improving the precision [28], [32], [41], starting with a pre-determined list of well-known people. Both are infeasible in real-world video archive scenarios. In this work, to our knowledge, we present the first method for automatically determining the usefulness of a set of search-results, before automatically removing outliers from the retrieved results of those deemed useful, resulting in a completely automatic, scalable and high-precision method.

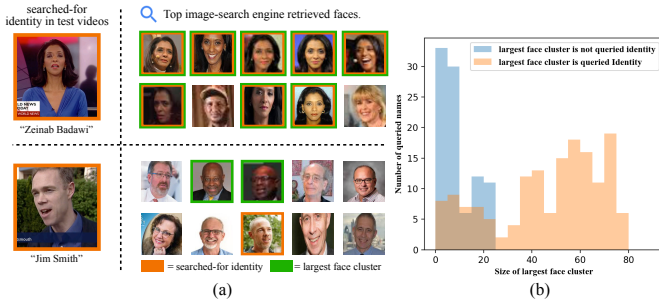


Fig. 3. (a) Examples of the top-10 retrieved image-search engine results for two different candidate names. For the “Zeinab Badawi” query (top row), the large face cluster (shown in green) has 8 faces, and all depict the queried person (shown in orange). The actual cluster for “Zeinab Badawi” in the top-100 retrieved images has 76 faces. For the “Jim Smith” query (bottom row), the largest face cluster is far smaller with 2 images, and does not depict the searched-for person. The observation is made that if the largest face cluster has many images, then it will likely contain the searched-for person. (b) A plot of whether the largest face cluster contains the searched-for identity for 350 randomly chosen candidate names found for the BBC Videos dataset. Clearly, if the largest face cluster has many faces in it (e.g. more than 30 in the top-100) then it is very likely to contain the searched-for person.

III. STAGE 1 – USING IMAGE-SEARCH ENGINES AS SOLE/CORROBORATING EVIDENCE

In this section, we describe our approach for using image-search engines to obtain sole or *corroborating evidence* for labelling people in videos. As a precursor, there are three cases for the usefulness of image-search engines depending on the candidate name: if the person is *famous*, many images of them will be found online (retrieved results will have high precision); for the *less-famous* people, there may be one or two images of them returned; and for the *never-famous* people there will be none (retrieved results have zero recall).

This stage of the method proceeds in two steps. The first is **obtaining candidate names**. The names of appearing people are not known *a-priori* and therefore must be first automatically gathered before any labelling can commence. For video material like TV news broadcasts or movies, we use three different sources: (i) IMDB names lists (ii) Scene text: names appearing in text displayed on screen during the programmes, such as overlaid banners. (iii) Speech: names appearing in the speech of the audio track. This step results in a set of candidate names for each episode. Details of the fully automatic processing to obtain these candidate names are given in Section V-A and the Supplementary Material.

The second step is **labelling famous and less-famous people**. Given a candidate name, we first determine if the person is *famous* or not using a novel method explained below based on downloaded images from an image-search engine. We then proceed in one of two ways: (1) For each *famous* name, a face model is built from the downloaded images and used as the sole evidence to label that person throughout the video; or (2) For the *less-famous* people, we use the temporal occurrence of their spoken or written (displayed) name in a video as primary evidence, and any single occurrence of them in retrieved image-search engine results as *corroborating evidence*. There is no labelling method for the *never-famous*

people, as image-search engines provide no examples of their appearance. The following describes these methods in detail.

Sourcing Candidate Names from IMDB: The name lists are freely and automatically obtained starting from just the name of the programme (which often can be found automatically in video metadata). These lists do not constitute curated cast lists as they often contain thousands of names of briefly appearing characters. This differs from previous non-scalable work on automated face labelling, e.g. [37], [40], [41], which use a curated list of *famous* appearing names.

Classifying Candidate Names as famous: When a candidate name is queried in the image-search engine, we use the key observation that if many of the top-ranked retrieved results correspond to the same person, then this person is *famous*. This observation is illustrated in Figure 3. In detail, faces are detected in the top 100 ranked results, and clustered using Agglomerative-Clustering [29] on their L2-normalised face embeddings [13] (using a cosine-distance threshold of 0.7). If the largest face-image cluster has more than α faces (in this work we use $\alpha = 30$, learnt on a validation set as described in Section VI-A, and also illustrated in Figure 3), then the identity is classified as *famous*.

Building a Face-Identity Model For the famous People:

For the candidate names that are classified as *famous*, we simply build a biometric model for that identity and use it to label any face-tracks depicting that person in all test videos, as shown in Figure 2. We take the largest cluster of face-embeddings from the downloaded images and average-pool and L2-normalise them into a single embedding. This single embedding is now a face-identity model that can be used for labelling. Taking only the embeddings from the largest cluster serves the purpose of removing false positives from the downloaded images. Face-tracks in the test videos are then labelled by measuring the cosine similarity between their embeddings and the face-identity embedding. If a face-track embedding has a similarity score higher than a threshold learnt on a validation set, then it is labelled with the famous name.

Finding Corroborating Evidence for less-famous People:

When a candidate name is not classified as *famous*, there may still be a few images of the person in the downloaded images (e.g. the bottom example in Figure 3). These low-precision images cannot serve as the sole evidence as was the case for the *famous* people, but can serve as *corroborating evidence*. Hence, for *less-famous* people we use the temporal occurrence of their spoken or written name as primary evidence of their appearance, and then a single correct retrieved face from image-search engines as the necessary *corroborating evidence* to label. In detail, the *corroborating evidence* is that at least one of the 20 top ranked faces from an image-search engine matches the face-track that appeared in the test video when the name was found. The face-track is then labelled with that name. We are here using 1-to-1 face verification. This is less accurate than the template-based (many-to-one) face verification that we use for the *famous* names, however seeing as the evidence is supported by the presence of the name in the scene or audio track, it is sufficient here.

IV. STAGE 2 – BOOSTING THE NUMBER OF TAGS

In this section, we describe the two methods used for boosting the number and variety of tags in the test videos. This includes fusing the evidence sources of face-appearance and speech, as well as query expansion.

A. Fusing face-appearance and voice

This section explains how we use additional information from the speech modality as corroborative evidence to label further face-tracks when the face-appearance alone is not enough to make a confident tag. For each tagged face-track, we use Active Speaker Detection [18], [19] (ASD), to classify whether the face is speaking. For the speaking faces of each tagged person, we extract temporally aggregated speaker embeddings [17] using the overlapping audio segments, which after average pooling form a speaker model for that identity. For the remaining un-labelled speaking face tracks in the video, we compute the similarity score between the speaker embedding and the speaker ID models (voice score), and the similarity score between face and the face ID models (face score). We then simply average the two scores (fusion score), and label the face-track if it is above a given threshold. We find empirically that the simple rule of averaging the two scores is highly effective. For any speaking face-track to be incorrectly classified with the fusion score, both the voice and the face score needs to be high for the same, incorrect identity. We see during experiments that this is very rarely the case, due to the lack of coupling between the modalities. In the supplementary material we present experiments with more complex architectures and show that this very simple rule achieves comparable performance. In the next sections we refer to this method as “Stage 2 Fusion”.

B. Query Expansion

Query expansion (QE) is a popular re-ranking method [2], [11], [16], [45]. Methods assume top ranking instances to be from the same class as the query, and use these to supplement the original query to create a new, superior ranking. Nagrani and Zisserman [37] perform QE by training a new classifier for each identity with their top ranked test-video tags, and show it to be helpful for crossing the domain gap between the search engine face-images and the TV-material face-tracks (similarly explored in [9] for voice). We adopt the same technique in this work, except we do not train any additional parameters at this stage, but instead just average-pool all tagged face-image embeddings to form a new face-identity model, which is then used to make further tags. In the next sections we refer to this method as “Stage 2 QE”.

V. DATASETS, EVALUATIONS AND IMPLEMENTATION

In this section we first describe the datasets and the evaluations used for assessing the method, and then give some details on implementation. The video labelling method can proceed completely automatically given only the programme names. This ‘plug and play’ automation is assessed in experiments on two datasets: BBC Videos and MediaEval. We also test

TABLE I
DATASET STATISTICS AND INFORMATION ON WHICH PARTS OF THE METHOD ARE TESTED ON EACH OF THE DATASETS. ANNOTATIONS ARE EITHER PROVIDED AT THE FACE-TRACK LEVEL (BBC VIDEOS, SHERLOCK) OR AT THE SCENE LEVEL (MEDIAEVAL).

Dataset	BBC Videos	MediaEval [44]	Sherlock [37]
No. Identities	66	1,971	31
No. Annotations	1,971	6,889	5,246
No. Videos	5	79	3
No. Hours	2.2	49.1	4
Test Stage 1	✓	✓	
Test Stage 2 - Fusion	✓	✓	✓
Test Stage 2 - QE	✓	✓	✓

out Stage 2 of the method on the standard person identification benchmark, Sherlock, where cast lists and corresponding face-images are provided. Statistics of the three datasets used are given in Table I. Different test protocols are used, when either testing the whole method, or just Stage 2, so that previous published methods can be compared to.

BBC Videos: This dataset consists of five episodes of different BBC television programmes (BBC news, BBC World News, Newsnight, Question Time). The challenging dataset for identification provides annotations for all *human-identifiable* characters, from road-side interviewees to well-known politicians. These are the people whose names are alluded to somewhere in the episode, or who are well-known *i.e.* which a human-annotator with access to the internet could annotate. This includes most people barring audience members, pedestrians, etc. Only the names of the programmes are provided to the method. The first episode constitutes the validation set.

MediaEval [44]: This dataset featured in the MediaEval 2015 challenge “Multimodal person discovery in broadcast TV” [44]. The test set [8] consists of 79 episodes from the French TV news show, “*Le 20 heures*”, with a total of 6,889 speaking faces annotated at the scene level. Only the name of the programme is provided to the method.

Sherlock [37]: This dataset consists of three episodes of the crime drama show “Sherlock”, where the face-tracks depicting main characters have been annotated. Each episode is approximately 80 minute long. A cast list and face-identity models (in the form of image-search engine images) are provided for each of the annotated characters, and the task is to classify each face-track by identity. The fast-paced show contains many visually challenging scenes (*e.g.* dark, quick camera movement), making it a difficult task for person labelling. For fair comparison to previous works, the face-tracks provided with the Sherlock dataset are used in the experiments.

Evaluation Metrics: When starting from the programme name alone, the BBC Videos and MediaEval experiments constitute open-set retrieval tasks, and so are evaluated using retrieval metrics: Precision, Recall, mAP, class recall – a measure of how many of the total classes have had 1 instance correctly retrieved. The dataset testing Stage 2 alone (Sherlock) provides a closed-set classification task, and so is evaluated using classification accuracy.

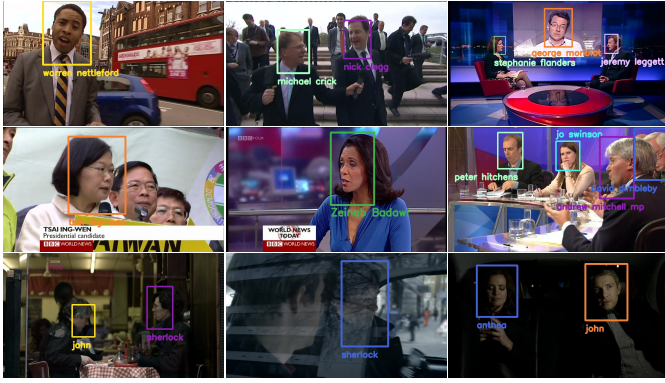


Fig. 4. Correctly labelled faces from BBC Videos (top, middle row) Sherlock (bottom row). These visually disparate datasets set challenging scenarios for person labelling, such as low resolution, lighting and extreme poses. BBC Videos: (left) Face labels obtained using *corroborating evidence* from image-search engines. (middle) Talking faces labelled by the Fusion step. (right) The QE step labels small, occluded faces, and extreme poses. Sherlock: From left to right, extreme poses, occlusion from glare, and dark faces.

A. Implementation Details

For automatic preparation of each dataset, we compute face-tracks (from face-detections [18]) and speech-turns (linked to faces using Active-Speaker Detection (ASD) [18] for each test-video. Additionally, we extract approximate transcripts using Automatic Speech Recognition [1] (ASR) and any scene-detected-text is found using Optical Character Recognition (OCR) techniques [22], [35]. Full details on all video pre-processing method are given in the suppl. material.

Face and speech embeddings. For face-tracks and speech-turns we use pre-trained embeddings to perform face [13] or speaker [17] verification, respectively. For face-tracks, embeddings from each face detection are average pooled and L2-normalised into a single embedding. For speech turns, temporal average pooling of the features along the time domain produces a single utterance-level embedding.

VI. RESULTS

In this section, we first investigate the automatic method of classifying whether a name is *famous* or not, and then evaluate either both stages, or just Stage 2, on the three datasets.

A. Determining if people are famous or not

Here we investigate the choice of the *famous* classification parameter α on candidate names for the BBC Videos dataset. A high α (> 25) means that it is likely that the faces in the largest cluster of downloaded images correspond to the searched-for person. This results in very confident face-identity models, and subsequent perfect face-track precision levels (Figure 5a) when these models make correct tags. However, a high α also means that the *famous* classification is limited to very well-known people. This results in low face-track recall (Figure 5b), as many of the candidate names are then not classified as *famous* and so are not tagged. A low α (< 25) leads to many names being classified as *famous*. This leads to a high face-track recall (> 0.85 for Stage 1, and > 0.93 for Stage 2) because face-identity models are built for

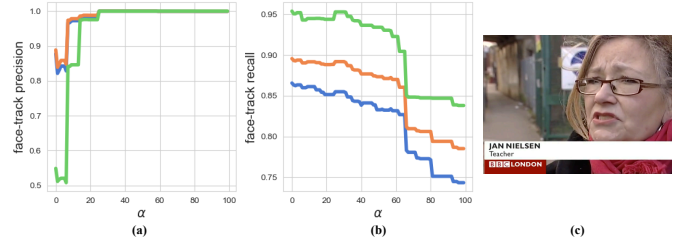


Fig. 5. Analysis of the *famous* threshold α on the BBC Videos dataset. (a) the precision of the face-track tags as α increases. (b) the face-track recall as α increases. The colors are: blue for Stage 1, orange for Stage 2 Fusion only, and green for Stage 2 Fusion + QE. (c) A missed tag from the BBC Videos dataset (see Section VI-B). This figure is optimally viewed in colour.

TABLE II
BBC VIDEOS DATASET RESULTS.

BBC Videos			
Method	Face-Tracks		Class Recall
	Precision	Recall	
Stage 1	1.0	0.860	0.91
Stage 1 + (Stage 2 Fusion only)	1.0	0.894	0.91
Stage 1 + (Stage 2 QE only)	1.0	0.934	0.91
Stage 1 + (Stage 2 Fusion + QE)	1.0	0.953	0.91

many people in the videos. However, Figure 3 shows how at low α , the largest cluster does not always depict the searched-for person, so face-identity models become polluted with false positives. This leads to poor face-track precision as incorrect tags are made with bad face-identity models. Stage 2 QE (blue in Figure 5a/b) then creates new face models from incorrect tags, worsening the problem (precision < 0.6 for $\alpha < 10$).

The chosen value of $\alpha = 30$ reflects the balance between achieving high face-track recall, whilst ensuring high precision. This value selects 1,967 *famous* names from a total of 2,906 sourced from IMDB. The IMDB names lists are not curated cast lists, and so can contain hundreds of names irrelevant to this task. For written and spoken names, 129 are classified as *famous* and 170 as *less-famous*. α is not influenced by the video being labelled and so remains constant.

B. BBC Videos

The BBC Videos dataset is used to test the full automated pipeline (Stages 1 and 2). This dataset was annotated exclusively for this research. This means that we cannot compare to prior work. Instead we use the dataset to perform a quantitative analysis of the different stages of the method. Results are shown in Table II. For each face-track either the correct label is assigned, the incorrect label is assigned, or we refuse to predict a label as no model has sufficient confidence.

Stage 1: The intended design choice is to only present correctly labelled face-tracks, so the classifier threshold (for tagging faces in the test videos) is chosen on the validation set such that we make no labelling mistakes. This results in a precision of 1.0 across all episodes, meaning that no people were incorrectly tagged, whilst achieving a high face-track recall of 0.86. A class recall of 0.91 indicates that 61 of the 66 people in the dataset were correctly tagged at least once (the missed classes are those written or spoken names for whom

no *corroborating evidence* could be found on search engines). These results are impressive given that the only information provided was the programme name. Most image-search engine images are frontal faces (see Figure 3), and this is reflected in the face tags made by this stage (see Figure 4, left column, top/middle row). As no further evidence of new, unlabelled identities is found after Stage 1, the Class Recall does not increase further. Further details are given in the suppl. material.

Stage 2: The fusion method improves face-track recall by 3.4%. Here, the voice modality is harnessed to confidently tag extreme face poses that Stage 1 could not (see Figure 4, central column, top/middle row). The QE step (Stage 1 + (Stage 2 QE only)) leads to a comparatively larger 6.4% improvement over Stage 1. QE here is able to bridge the domain gap from image-search engine images to the test video, and increase the number of correct tags. For the combined experiment (Stage 1 + (Stage 2 Fusion + QE), QE builds new identity models from the tags made by both Stage 1 and the fusion step. When combined, these offer a rich variety of poses, that are representative of the within-class variations. This therefore leads to the largest improvement to 0.953 face-track recall, where faces in a range of poses are tagged throughout the videos (see Figure 4, right column, top/middle row).

Figure 5c shows an example of a missed tag. This person was not labelled even though their name is displayed, as no *corroborating evidence* could be found on search engines of their appearance. This problem is non-trivial, as a displayed name does not always correspond to the displayed person. Human annotators used the fact that this person is introduced in a prior scene where they were not present. For an automated process, this tag requires complex, longer term reasoning capabilities. This opens possibilities for future work.

C. MediaEval

Our results on the MediaEval dataset are shown in Table IIIa. We experiment with both the original MediaEval 2015 challenge rules [44] (no external biometric models may be used, so no image-search engines), and with the combined Stages 1 and 2 of the proposed method. Challenge participants use the strong prior that a written name is very likely to belong to the co-occurring speaking-and-visible face in the scene. Impressively our MediaEval 2015 rules method gains a 10% improvement on the original baseline method [8], and a significant 2.2% improvement on the state-of-the-art [43], while using just pre-trained out-of-the-box features. Using our full method (Stages 1 and 2), results in an impressive further 3.88% improvement in mAP. This gain is seen because our method is able to provide labels regardless of the proximity of a face to their corresponding written name, and is able to use stronger biometric models through the assistance of image-search engines. Our MediaEval 2015 rules method, as well as the previous baselines, fails to identify very well known people if their name is never spoken or found written. Without any manual supervision, our proposed method is able to correctly identify these people.

TABLE III
(A) RESULTS ON THE MEDIAEVAL DATASET. (B) RESULTS ON THE SHERLOCK DATASET - VALUES ARE THE PER-CHARACTER CLASSIFICATION ACCURACY FOR THE MAIN CHARACTERS IN EACH OF THE THREE EPISODES. OUR METHOD IMPROVES OVER THE STATE-OF-THE-ART FOR BOTH DATASETS. KEY: ME: MEDIAEVAL.

MediaEval [44]		Sherlock [37]			
Method	mAP (%)	Method	E01	E02	E03
Baseline [8]	74.89	Nagrani & Zisserman [37]	0.92	0.90	0.88
SOTA [44]	82.80	Ours (face)	0.88	0.81	0.86
Ours (ME 2015 rules)	85.22	Ours (face + Stage 2)	0.95	0.93	0.94
Ours (Stage 1 + Stage 2)	89.10				

D. Sherlock

The results for testing out Stage 2 of our method on the Sherlock dataset are shown in Table IIIb. The dataset provides face images from image-search engines for each identity in the test set, so it is not necessary to run the complete Stage 1 of our method. Instead we obtain the Stage 1 labels as follows: face embeddings are extracted for each of the provided images, and average pooled for each identity to give a face model; then the cosine similarity between each of the identity models and the test video face-tracks is computed; finally, each track is labelled with the identity and score of the model that has the maximum cosine similarity. This gives a preliminary set of face labels ('Face' result in Table IIIb), from which Stage 2 can now boost tag numbers ('Face + Stage 2' result in Table IIIb). Our method surpasses the previous state-of-the-art [37] considerably by a margin of 3-6% on all three episodes. This is particularly impressive, as the original work uses extra parameters for a SVM multi-way classifier, whereas our work simply uses a nearest neighbour classifier from aggregated identity features, which has no extra parameters and requires no extra training. The improvements are due to superior face embeddings, and also to the evidence fusion, which is able to classify characters where the original work, which treated modalities independently, could not. Figure 4 show examples of tagged faces on this challenging dataset.

VII. CONCLUSIONS

In this paper, we propose a novel method for the automated labelling of people in videos through the use of *corroborating evidence*, both from image-search engines, and from different information modalities. The method performs impressively over a set of visually disparate domains both when starting from just the programme name, and also when testing out certain stages, setting new state-of-the-art results in the process. This method therefore provides a robust and reliable technique for labelling large video archives.

ACKNOWLEDGMENT

This work is supported by a EPSRC DTA Studentship, and the EPSRC programme grant Seebibyte EP/M013774/1. We are grateful to Arhsa Nagrani, Shaya Ghadimi, and Maya Gulieva for proof reading, and the reviewers for their helpful feedback.

REFERENCES

- [1] T. Afouras, J. S. Chung, and A. Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," in *Proc. ICASSP*, 2020.
- [2] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. CVPR*, 2012.
- [3] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," in *Proc. ACCV*, 2020.
- [4] M. Bauml, M. Tapaswi, and R. Stiefelhof, "Semi-supervised learning with constraints for person identification in multimedia data," in *Proc. CVPR*, 2013.
- [5] M. Bendris, B. Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in tv content using audio-visual sources," in *Proc. CBMI*, 2013.
- [6] T. L. Berg and D. A. Forsyth, "Animals on the web," in *Proc. CVPR*, 2006.
- [7] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in *Proc. ICCV*, 2013.
- [8] H. Bredin, "Mediaeval person discovery task, evaluation package 2015," <https://github.com/MediaevalPersonDiscoveryTask/EvaluationPackage2015>, 2015.
- [9] A. Brown, J. Huh, A. Nagrani, J. S. Chung, and A. Zisserman, "Playing a part: Speaker verification at the movies," in *Proc. ICASSP*, 2021.
- [10] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman, "Smooth-AP: Smoothing the path towards large-scale image retrieval," in *Proc. ECCV*, 2020.
- [11] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic query expansion using SMART," in *TREC-3 Proc.*, 1995.
- [12] L. Canseco, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.
- [13] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.
- [14] K. Chatfield and A. Zisserman, "Visor: Towards on-the-fly large-scale object category retrieval," in *Proc. ACCV*, ser. Lecture Notes in Computer Science. Springer, 2012.
- [15] Z. Chen, W. Zhang, B. Deng, H. Xie, and X. Gu, "Name-face association with web facial image supervision," *Multimedia Systems*, 2019.
- [16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. ICCV*, 2007.
- [17] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *Proc. INTERSPEECH*, 2020.
- [18] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading*, ACCV, 2016.
- [19] S. Chung, J. S. Chung, and H. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3965–3969.
- [20] R. G. Cinbis, J. J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in TV video," in *Proc. ICCV*, 2011.
- [21] T. Cour, B. Sapp, A. Nagle, and B. Taskar, "Talking pictures: Temporal grouping and dialog-supervised person recognition," in *Proc. CVPR*, 2010.
- [22] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. AAAI*, 2018.
- [23] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! My name is... Buffy' – automatic naming of characters in TV video," in *Proc. BMVC*, 2006.
- [24] —, "Taking the bite out of automatic naming of characters in TV video," *Image and Vision Computing*, 2009.
- [25] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proc. ICCV*, 2005.
- [26] P. Gay, G. Dupuy, C. Lailler, J.-M. Odobez, S. Meignier, and P. Deléglise, "Comparison of two methods for unsupervised person identification in tv shows," in *Proc. CBMI*, 2014.
- [27] M.-L. Haurilet, M. Tapaswi, Z. Al-Halah, and R. Stiefelhof, "Naming tv characters by watching and analyzing dialogs," in *Proc. WACV*, 2016.
- [28] A. Holub, P. Moreels, and P. Perona, "Unsupervised clustering for google searches of celebrity images," in *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2008.
- [29] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [30] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin, "Automatic named identification of speakers using diarization and asr systems," in *Proc. ICASSP*, 2009.
- [31] M. Köstinger, P. Wohlhart, P. Roth, and H. Bischof, "Learning to recognize faces from videos and weakly related information cues," in *avss*, 2011.
- [32] D.-D. Le and S. Satoh, "Unsupervised face annotation by mining the web," in *ICDM*, 2008.
- [33] N. Le, H. Bredin, G. Sargent, M. India, P. Lopez-Otero, C. Barras, C. Guinaudeau, G. Gravier, G. B. da Fonseca, I. L. Freire *et al.*, "Towards large scale multimedia indexing: A case study on person discovery in broadcast news," in *Proc. International Workshop on Content-Based Multimedia Indexing*, 2017.
- [34] W.-H. Lin, R. Jin, and A. Hauptmann, "Web image retrieval re-ranking with relevance model," in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, 2003.
- [35] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Proc. ECCV*, 2018.
- [36] J. Maclair, S. Meignier, and Y. Estève, "Speaker diarization: about whom the speaker is talking?" in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, 2006.
- [37] A. Nagrani and A. Zisserman, "From Benedict Cumberbatch to Sherlock Holmes: Character identification in TV series without a script," in *Proc. BMVC*, 2017.
- [38] O. M. Parkhi, E. Rahtu, Q. Cao, and A. Zisserman, "Automated video face labelling for films and tv material," *IEEE PAMI*, 2020.
- [39] O. M. Parkhi, E. Rahtu, and A. Zisserman, "It's in the bag: Stronger supervision for automated face labelling," in *ICCV Workshop: Describing and Understanding Video & The Large Scale Movie Description Challenge*. IEEE, 2015.
- [40] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "On-the-fly specific person retrieval," in *International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, 2012.
- [41] —, "Deep face recognition," in *Proc. BMVC*, 2015.
- [42] P. T. Pham, M.-F. Moens, and T. Tuytelaars, "Naming persons in news video with label propagation," in *ICME*, 2010.
- [43] J. Poignant, L. Besacier, and G. Quénot, "Unsupervised speaker identification in tv broadcast based on written names," *ACM Transactions on Audio, Speech, and Language Processing*, 2014.
- [44] J. Poignant, H. Bredin, and C. Barras, "Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015," *Multimedia Tools and Applications*, 2017.
- [45] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *Journal of the American Society for Information Science*, 1999.
- [46] S. Satoh and T. Kanade, "Name-it: Association of face and name in video," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.
- [47] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *Proc. ICCV*, 2007.
- [48] J. Sivic, M. Everingham, and A. Zisserman, "'Who are you?' – learning person specific classifiers from video," in *Proc. CVPR*, 2009.
- [49] M. Tapaswi, M. Bauml, and R. Stiefelhof, "'knock! knock! who is it?' probabilistic person identification in tv series," in *Proc. CVPR*, 2012.
- [50] Y. Tian, L. Zhou, Y. Zhang, T. Zhang, and W. Fan, "Deep cross-modal face naming for people news retrieval," *Proc. TKDE*, 2019.
- [51] P. Wohlhart, M. Köstinger, P. M. Roth, and H. Bischof, "Multiple instance boosting for face recognition in videos," in *DAGM-Symposium*, 2011.
- [52] J. Yang, R. Yan, and A. G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in *Proc. ACMM*, 2005.