From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script

Arsha Nagrani
arsha@robots.ox.ac.uk/
Andrew Zisserman
az@robots.ox.ac.uk/

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

Abstract

The goal of this paper is the automatic identification of characters in TV and feature film material. In contrast to standard approaches to this task, which rely on the weak supervision afforded by transcripts and subtitles, we propose a new method requiring only a cast list. This list is used to obtain images of actors from freely available sources on the web, providing a form of partial supervision for this task. In using images of *actors* to recognize *characters*, we make the following three contributions: (i) We demonstrate that an automated semi-supervised learning approach is able to adapt from the actor's face to the character's face, including the face *context* of the hair; (ii) By building voice models for every character, we provide a bridge between frontal faces (for which there is plenty of actor-level supervision) and profile (for which there is very little or none); and (iii) by combining face context and speaker identification, we are able to identify characters with partially occluded faces and extreme facial poses.

Results are presented on the TV series 'Sherlock' and the feature film 'Casablanca'. We achieve the state-of-the-art on the Casablanca benchmark, surpassing previous methods that have used the stronger supervision available from transcripts.

1 Introduction

One of the long-term objectives in computer vision is video understanding, for which an important step is the ability to recognize peoples' identities under unconstrained conditions. In this area, TV dramas, sit-coms and feature films have provided a strong test bed, where the goal has been to recognize characters in video using their faces (albeit usually only frontal faces). Achieving this goal is crucial to story understanding, and perhaps more directly, to generate metadata for indexing and intelligent fast forwards (e.g. "fast forward to scenes when Sherlock speaks to John"). Since the work of Everingham *et al.* [11], most methods have made use of a transcript aligned with the subtitles to provide weak supervision for the task [2, 7, 8, 9, 10, 15, 19, 23, 27, 28, 29].

In this paper, our goal is also to recognize characters from video material; the novelty of our approach is that we eschew the use of transcripts and subtitles in favor of a simpler solution that requires only the cast list. The key idea we explore, is that it is possible to get *partial* supervision from images of the actor's faces that are easily available using web image search

engines such as Google or Bing's Image Search. This supervision is only partial, however, because of the disparity between appearances of actors and their corresponding characters. Images of actors available online are typically from red-carpet photos or interviews, whereas a character can have a substantially different hair style, make-up, viewpoint (extreme profile, rather than the near frontal of actor interviews), and possibly even some facial prosthetics. In essence, the problem is one of domain adaptation, as shown in Figure 1.

Using only actor-level supervision to recognize characters, we make the following contributions: (i) We demonstrate that an automated semi-supervised learning approach is able to adapt from the actor's face to the character's face, including the face *context* of the hair; (ii) We use active speaker detection and speaker identification to build voice models for every character, and show that such models can act as a bridge between frontal faces (for which there is plenty of actor-level supervision) and profile (for which there is very little or none); and (iii) by combining face context and speaker identification, we are able to identify characters with partially occluded faces and further profile faces. The outcome is a set of characters labelled in both frontal and profile viewpoints.

We demonstrate the method on two datasets: the TV series 'Sherlock', episodes 1–3, which will be used as the running example in this paper; and the feature film 'Casablanca' which is used as a benchmark for this task. On Casablanca, we achieve state-of-the-art results, surpassing previous methods that have used the stronger supervision available from transcripts and subtitles.

We pursue a transcript-free approach for a number of reasons. Transcripts, typically obtained from IMDB or fan sites, are difficult to find for all films or entire seasons of a TV series. They also tend to come in a variety of different styles or formats, ensuring much work must go into standardizing formats. In contrast, cast lists are easy to obtain for all episodes of a show from IMDB. Cast lists can also be obtained visually from the credits (using OCR) and hence no external annotation is required. We can also aim to label every credited character, and not just principal characters or those with speaking parts (as is typically done in approaches using transcripts).

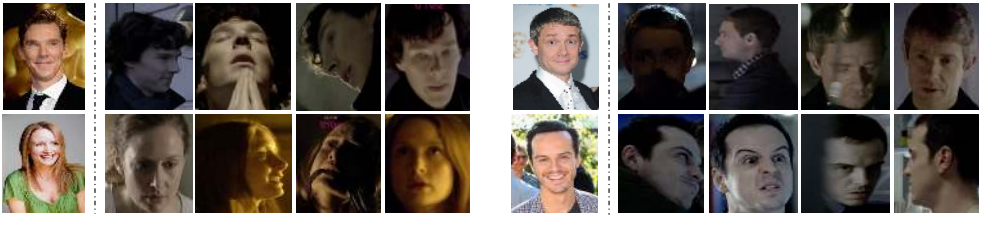

Figure 1: Examples of occurrences of actors in source (leftmost image in each set) and target domains for four identities. The source images (from web images) can differ from the target images (from the TV material) in hairstyle, makeup, lighting, and viewpoint.

1.1 Related Work

When weak supervision is available from transcripts, then the problem can be cast as one of ambiguous labelling [8] or Multiple Instance Learning (MIL), as employed by [2, 15, 29, 30].

In our case, where the partial supervision is provided by web images given the actor's name, the approach is related to the on-the-fly retrieval of [18], although previous work has not considered the domain adaptation aspect developed here. In order to fully solve the domain adaptation problem, we leverage the use of the audio track to classify speech segments extracted automatically from the video. While speaker detection is used to obtain

labels from transcripts and subtitles, to the best of our knowledge very few works [26, 28] have attempted to build voice models to aid automatic labelling of tracks. By propagating labels from one modality to another (facetracks to audio segments and vice versa), we use a co-training approach similar to that outlined in [1].

Most previous methods for automated character labelling have concentrated on frontal faces, with only [2, 19, 23, 27, 28] going beyond frontal faces to also consider profiles. [31] used poselet-level person recognizers in order to deal with large variations in pose, while [22] explicitly employed the hair as a measure for tracking and was thus able to extend some frontal tracks to profiles, though did not detect and track purely profile faces. The consequence of this is that the non-profile methods ignore (and thus do not attempt to identify) characters that appear in profile views. Those that have attempted to also identify profile views [2, 19, 23, 27, 28], have often found inferior performance in profile, compared to frontal, due to an inability to learn profile face classifiers when primarily frontal faces are available for supervision. We show that voice models can be used to overcome this problem.

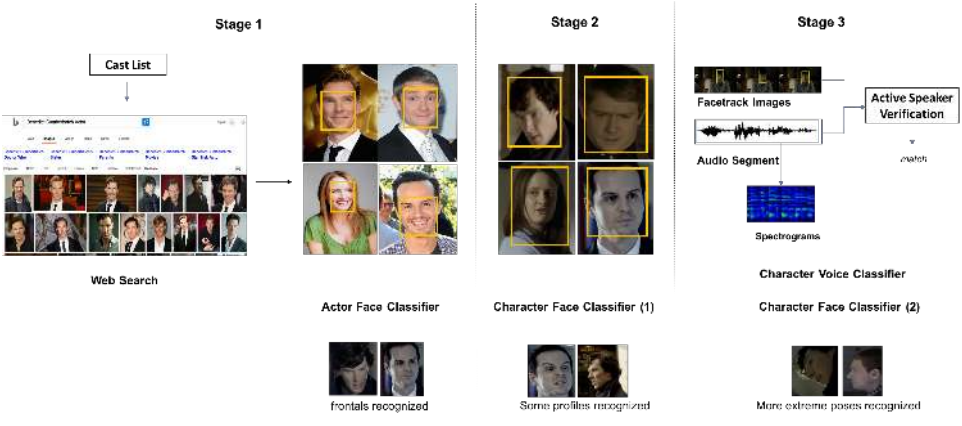

Figure 2: Overview of the three stages of the approach.

2 Overview

As is standard in almost all previous approaches for this task [2, 7, 8, 9, 10, 15, 19, 27, 28, 29], the basic unit of labelling for our task is a face track extracted from the video. Starting from images of the actors available online, we build a series of multiclass SVM classifiers, each relying on the output of the previous classifier for supervision. Using this semi-supervised technique, we can achieve robustness to pose, lighting, expression and partial occlusions with very little initial supervision.

Our goal is to label all the face tracks (both of frontal and profile faces) for the actors/characters given in the cast list. The approach proceeds in three main stages (shown in Figure 2):

1. Actor face classifier: The first stage is to obtain initial labels for the face tracks from actor images alone. Starting from the cast list, images of the actor's faces are collected from search engine results available on the web (Section 3.1). A facial classifier is then trained for each actor from these images, and used to classify the face tracks. As demonstrated in Figure 1, actor images have fundamental differences in pose, lighting, expression and make up to the target images, and consequently the face tracks labelled correctly at this stage are mostly limited to the relatively easy frontal images.

2. Character face classifier: The second stage involves building face classifiers for each *character* (Section 3.2). Since some initial labels are available from stage 1 (from confident actor face classifiers), we can now train this character classifier on images from the video directly. This allows us to learn the face context (including the hair) of the character. The face tracks that the actor face classifier is not confident for, are then reclassified with these character face classifiers. The outcome is that some profile and partially occluded faces are now recognized correctly.
3. Character voice and face classifiers: The final stage involves adding the speech modality to improve results further (Section 4). First, face tracks that are speaking are determined using audio-visual correlation. Second, voice classifiers are built for each character using the face labels provided by stage 2 (once again, labels are propagated from *confident* character face classifiers). The face tracks for which the character face classifier is not confident are then reclassified by their voice. This corrects the labels of some of the profile face tracks corresponding to speaking characters. A final character face classifier is then trained including these newly corrected labels, and all tracks reclassified by this voice augmented character face classifier. In practice, profile faces not detected in stage 2, as well as those corresponding to speaking characters with very little actor level supervision, are identified in this stage.

3 Face Classifiers

We learn three different face classifiers. The first (actor face classifier) is trained on images obtained from the web and generates an initial classification of the face tracks. These classified face tracks are then used in a second round of training to learn the character's appearance (character face classifier 1). Finally, additional supervision is obtained from the speaker classifier, and the character classifier retrained (character face classifier 2).

3.1 Actor Face Supervision

A dataset of actor images is obtained automatically using the following steps:
1. Cast lists: Cast lists are obtained for every episode from the Internet Movie Data Base (IMDb). No formatting is required.
2. Image collection for each identity: Each actor name is then queried in Bing Image Search. Search results are improved by appending the keywords 'actor' and 'Sherlock TV series' to the name of each actor (this avoids erroneous actors with the same name). Exact duplicate images are automatically removed. Some amount of data is also obtained automatically from animated GIFs created by fans, providing up to 10 images per GIF.
3. Final manual filtering: This is for two reasons. First, there are erroneous images (corresponding to the wrong identity), these can trivially be removed (and in fact this could be automated). However, the second reason is that in order for the power of the method to be scientifically assessed, images of actors online which have been taken from the TV show, e.g. images of Benedict Cumberbatch where he is playing the character 'Sherlock' must be removed – this is purely to create a complete separation of domains (actor – character). In practice, this is only required for the more obscure actors, and is a quick and easy process with not more than 20 minutes of active human effort required for each episode.
4. Augmentation: Unlike disparities such as facial pose, occlusions and hairstyle, differences between the two domains such as illumination and resolution can be readily resolved using standard data augmentation techniques. Since images available online are usually brighter and of higher quality, the dataset is augmented in three different ways. Every image has a low contrast version, created using contrast stretching (normalization) with limits 0.4

and 1.0, a downsampled version obtained using bicubic interpolation (in an attempt to mimic lower resolution), and a horizontally flipped version. This makes the dataset 4 times larger.

3.2 Face Context for Character Face Classifier

Due to the domain adaptation issues discussed above, we aim to learn the face context of the character, but not the actor. To achieve this goal we distinguish between capturing only *internal* features (eyes, nose and mouth) and the extension to *external* features (contour, forehead, hair) by using facial bounding boxes of different extents. In the face classification literature, particularly for surveillance, internal features are found to be more reliable, since external information is more variable [16]; similarly, faces in the target domain (TV material) have different hairstyles and poses from those in the source domain (web images). Thus, when moving from source to target domain, it is sensible to restrict the support region for learning the actor face classifiers to internal features. Within the target domain, however, these external features prove essential in making the shift from frontal to profile faces, as well as providing a more robust model for cases where the face is partially occluded. The different support regions are shown in the stages 1 and 2 panels of Figure 2.

3.3 Face Track Classification

A similar process is used for each of the three face classifiers. Features vectors are obtained for each face track using a CNN (see implementation details Section 5). A linear SVM classifier is trained for each character, with the images of all other characters as negative training examples (one-vs-rest approach). The tracks are then ranked according to the maximum SVM score, and tracks above a certain rank threshold are assumed to have correct labels.

4 Speech Modality

While many approaches to this task focus on recognizing faces, very few leverage the use of the audio track to assist the automatic labelling of characters in TV material. As humans, this is an important cue for recognizing characters while watching TV; particularly if different characters have distinctive accents or pitch. Inspired by the success of speaker identification methods on TV broadcast material [12], we aim to create voice models for every character. Besides aiding in character identification, such models can be used in conjuction with appearance models for 'reanimating' or creating virtual avatars of popular TV characters [3]. Unlike news broadcasts, this task is more difficult in TV series for the following reasons: first, segmenting the audio track of a show or film into regions containing only a single speaker (speaker diarization) is difficult due to rapid shot changes, quick exchanges and crosstalk (common due to the conversational nature of speech); and, second, the speech segments themselves are degraded with a variety of background effects, crosstalk, music and laughter tracks. We tackle both these issues in the following two subsections.

1. Speaker diarization. Speaker diarization involves segmenting an audio stream into speaker homogeneous regions, which we henceforth term 'speech segments'. Speech segments contain speech from a single character, analogous to the face tracks used above. While [28] impose the constraint that the speaker is visible in the given shot, this assumption can be erroneous in the case of 'reaction shots', as well as voiceovers, dubbing and flashbacks. In order to obtain speech segments, we first extract audio segments corresponding to a single face track, and then apply Active Speaker Verification (ASV) to determine if the character in question is speaking. Since the ASV process is not precise to the frame level,

we only use face tracks where the ASV verifies that the face speaks for the entire track. The implementation details are given in Section 5.

2. Speaker identification. Once the speech segments are obtained, speaker identification is used to build voice models for every speaking character. While most recent approaches for speaker identification in TV [4, 21] use Gaussian Mixture Models (GMMs) trained on low dimensional features such as MFCCs [24, 25], we use here CNNs trained on raw spectrograms.

Since the number of speech segments extracted per episode is small (see Table 1), it is not possible to train a CNN from scratch. Hence we adopt a feature approach similar to that used for faces. We use the VoxCeleb CNN which is pretrained on a large speaker identification dataset [17] with raw spectrograms as input. This ensures that very little preprocessing is required (see implementation details Section 5). Using this network, we extract CNN based feature vectors and train a SVM classifier similar to the classifiers used above.
An important point to note here is that our method requires only speaker identification (and not speech recognition), and hence works for videos filmed in any language.

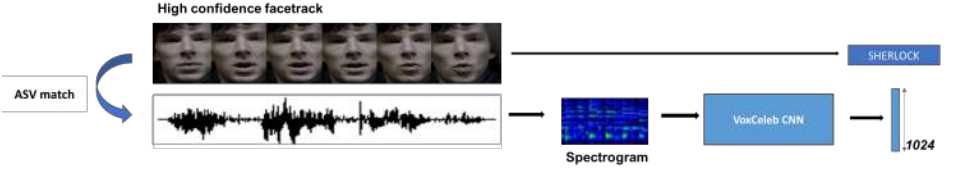

Figure 3: Speaker Identification. The entire speech segment spectrogram is represented by a 1024-D vector, and labels are propagated from the high confidence facetracks.

5 Implementation Details

5.1 Face

Face tracking: Face tracking employs the 'tracking by detection' approach used by [11]. This involves five major steps: (i) frame extraction, (ii) shot boundary detection, (iii) face detection, (iv) face tracking and (v) face/non-face classification. While shot boundaries are detected by comparing colour histograms across consecutive frames, face detection is done using a faster R-CNN model trained on the large scale WIDER face dataset [14]. Within each detected shot, face detections are grouped together into face tracks using a Kanade-Lucas-Tomasi (KLT) tracker. While most false (non-face) detections are filtered out during this tracking stage, some false positives persist. These are rejected (at the 'track' level) using a face/non face classifier based on the ResNet-50 [13] CNN architecture pre-trained on the VGG Face Dataset [20], and fine-tuned for the face/non-face task.
Face track Descriptors: Since the faster R-CNN detector used is trained to locate only the internal features, within the target domain the detection bounding boxes are extended by a factor of 0.25 to capture external features as well. Face track descriptors are then obtained following the approach of [20]. Detected face images are rescaled and passed through the VGG-Face CNN [20], and the output of the penultimate layer (the last fully connected) is computed to give 4096 dimensional feature descriptors. Feature vectors for every frame in the track are then sum-pooled and L2 normalized, giving a final 4096 dimensional vector for each track.

5.2 Voice

Active Speaker Verification (ASV): Face tracks with less than 50 frames (shorter than 2 seconds) are discarded as non speaking tracks. ASV is then performed using the two-stream CNN architecture described in [6], to estimate the correlation between the audio track and the mouth motion; albeit trained with a slight modification. Instead of using just cropped images of the lips, the entire face is used. This allows ASV even for extreme poses. A 1D median filter is then applied to obtain a confidence score for the track and to classify the corresponding audio segment as a speech or non-speech segment.

Audio preprocessing: Each speech segment is converted into a spectrogram using a hamming window of width 25ms, step 10ms and 1024-point FFT. This gives spectrograms of size $512 \times N$, where $N = 100L$, and L is the length of the speech segment in seconds.

CNN based descriptors: Spectrograms are then passed through the VoxCeleb CNN [17]. This is based on the VGG-M [5] architecture, modified to have average pooling before the last two fully connected layers, as well as smaller filter sizes to adapt to the spectrogram input. Features are extracted from the last fully connected layer (fc8), and since the entire speech segment can be evaluated at once by changing the size of the pool6 layer, a single feature vector of size 1024 is obtained for every speech segment (illustrated in Figure 3).

5.3 Face track Classification

After a model (face or voice classifier) is evaluated on the face tracks, the tracks are then ranked according to the maximum SVM score, $\max_y w_y^T x_i + b_y$, where y runs over the actor/character classes, and tracks above a certain rank threshold are assumed to have correct labels. These are then added to the training set of the next model.

Face: The rank threshold for the face classifiers is set manually for episode 1 at 50% of ranked face tracks, and then applied to all other episodes.

Speaker identification: All speech segments are split into train, test and val sets as follows: The train and val segments are those corresponding to confident face tracks with labels propagated from the face stage, and test segments are those tracks with low face confidence. The speaker classifier is extremely accurate, with almost 85% of test segments identified correctly (ground truths for this evaluation are obtained manually, however are not used in the pipeline). The face tracks corresponding to the top 80% of test speech segments are then 'corrected' with the speaker results.

6 Datasets

This section describes the two datasets used to evaluate the method.

Sherlock: All 3 episodes of the first season of the TV series "Sherlock". Each episode is about 80 minutes in length. Being a fast paced crime drama, this is a challenging dataset because it contains (i) a variety of indoor/outdoor scenes, lighting variations and rapid shot changes; (ii) variations in camera angle and position leading to unusual viewpoints, occlusion and extreme facial poses; and (iii) a large and unique cast as opposed to popular datasets for this task such as 'Big Bang Theory' (which are limited to 6 principal characters).

Casablanca: Casablanca is a full-length feature film used by Bojanowski *et al.* [2] for joint face and action labelling. Unlike 'Sherlock', the video is black and white. We use face tracks and annotations provided on the author's website [2] for evaluation.

Results on both datasets are given in Section 7, and statistics are summarized in Table 1. The average number of web images for each actor is less than 100, with more images

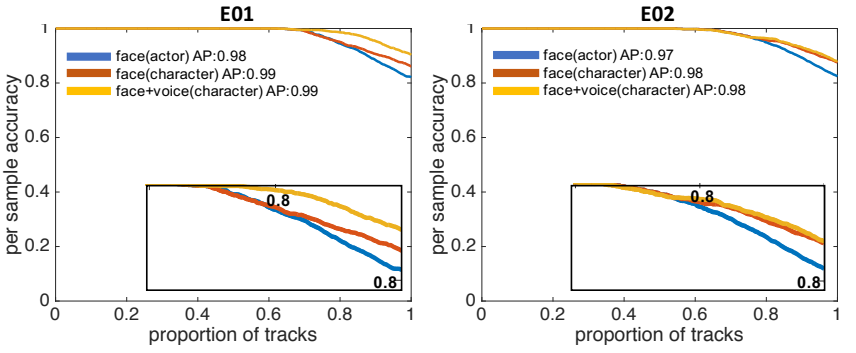

Figure 4: Precision-Recall Curves for the first 2 episodes of Sherlock. The inset shows a zoomed version of the top right corner.

available for the more famous actors (Benedict Cumberbatch: 147, Ingrid Bergman: 143). Furthermore, there is a gender dependent disparity, with more photos of female actors in circulation, a likely result of a higher propensity for photoshoots/modelling. For Casablanca, the images of the actors downloaded include both greyscale and colored images.

Episode	SE1	SE2	SE3	Casablanca
# Actor images	147 / 91.9 / 14	147 / 94.3 / 8	147 / 85.2 / 13	143 / 41.8 / 5
# Face tracks	1,731	1,775	1,740	1,273
# Speech segments	348	345	274	177
# Characters	15	12	16	17

Table 1: Statistics for episodes 1–3 of Sherlock, and for the film Casablanca. Where there are three entries in a field, numbers refer to the maximum / average / minimum.

7 Experiments

We evaluate the approach on both datasets, and measure performance using face track classification accuracy and Average Precision (AP).

Model	E1		E2		E3	
Face (actor)	0.85	0.98	0.82	0.97	0.82	0.97
Face (character context)	0.88	0.99	0.88	0.98	0.85	0.97
Face + Speech (character context)	**0.92**	**0.99**	**0.90**	**0.98**	**0.88**	**0.98**

Table 2: Classification accuracy (left) and average precision (right) for episodes 1-3 of Sherlock. Accuracy is the more sensitive measure.

7.1 Results on Sherlock

Results for all three face models are given in Table 2 and Figures 4 and 5, with example labelling shown in Figure 6. Overall, the performance is very high, with average precision for episode 1 almost saturated at 0.993.

Effect of character face models: As can be seen from Table 2, using the character face model can provide up to a 5% classification accuracy boost. This is particularly important for those actors where the hairstyles of their character differ greatly from those of the actor (see Figure 1).

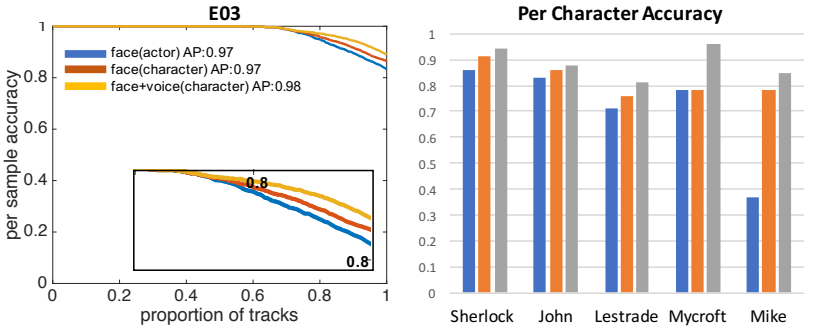

Figure 5: Precision-Recall Curve for episode 3 of Sherlock (left) and the corresponding character breakdown for this episode (right). Note the large performance increase with the character and voice models for Mike, who has little actor level supervision.

Effect of voice models: Voice identification also works very well. It increases performance for two main reasons: (1) it helps with the recognition of characters that have large speaking roles played by lesser known actors, e.g. the characters Mike and Mycroft (See Figure 5, right); (2) it helps with the identification of profile tracks of *all* characters for which actor-level profile supervision is severely lacking. This can clearly be seen in Figure 6 (row 2), which shows (from left to right) examples of a profile face, extreme pose face, and face with with little actor level supervision.

Error cases include heavy occlusions and small dark faces. As seen from Figures 4 and 5, all the errors occur at the high recall end of the curve where the classifier scores are low. This makes it possible to apply a 'refuse to predict' threshold, to avoid misclassification.

Figure 6: Examples of correct labelling results. From left to right: a blurred face in the background (Donovan), a dark frontal face and a partially occluded face (top row); profiles, extreme pose, and a lesser known actor (bottom row). Note: the original frames have been cropped for optimum viewing.

7.2 Comparison with the state-of-the-art

As noted by [2], feature films provide different challenges to TV series, and hence we modify our approach on the 'Casablanca' dataset as follows:

Actor level supervision: Images of actors available online are both greyscale and in colour.

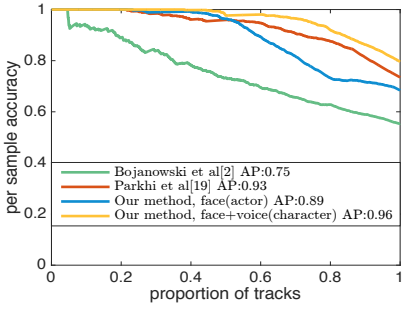

Figure 7: PR curve for Casablanca.

Model	AP
Sivic et al [27], as reported by [2]	0.63
Cour et al [10], as reported by [2]	0.63
Bojanowski et al [2]	0.75
Parkhi et al [19]	0.93
Our method	**0.96**

Table 3: Comparison to state-of-the-art.

Hence in the data augmentation step, we convert all images to greyscale, to match face tracks from the film.

Background characters: Films usually contain both principal characters (who appear often and have speaking parts), and background characters, who may make a single appearance and do not speak. While Sherlock has very few background characters, this is not the case for a feature film like Casablanca. In the many crowded bar scenes or outside on the streets, there are multitudes of background actors, with more than 30% of the tracks belong to the background category. To deal with this problem, we adopt the background classifier of [19], a simple SVM based classifier trained on track level statistics. Using this classifier, around 20% of the background tracks are identified. Tracks classified as background are then excluded in stage 1.

Results: As is clear from Table 3, we beat all previous attempts on this dataset, even though they use the stronger supervision provided by a combination of scripts and subtitles, and [19] also used VGG-Face CNN features. The performance boosts are extremely large for the principal speaking characters due to voice modelling. Most mistakes are for the background characters – these are very difficult to identify, largely because they are not credited and do not speak. Adding the ability to successfully model background characters will be an extension of this work.

8 Summary and Extensions

We propose and implement a simple yet elegant approach for automatic character identification in video, achieving high performance on a new dataset, 'Sherlock', and exceeding all previous results on the dataset 'Casablanca'. While face and voice models have been successfully created for each character, extensions to this work include training face and speech descriptors in a joint framework to create combined features. The speech modality can also be enhanced by applying ASV precise to the frame (and not track) level, in order to obtain more speech segments.

By eliminating the need for subtitles, transcripts and manual annotation, our approach is truly scalable to larger video datasets in different languages. We hope that the strength of our results will encourage the community to move towards more transcript-free approaches in the future.

Acknowledgements. We are grateful to Qiong Cao and Joon Son Chung for their help with this research. Funding was provided by the EPSRC Programme Grant Seebibyte EP/M013774/1.

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[2] P. Bojanowski, F. Bach, , I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proceedings of the International Conference on Computer Vision*, 2013.

[3] J. Charles, D. Magee, and D. Hogg. Virtual immortality: Reanimating characters from tv shows. In *Computer Vision - ECCV 2016 Workshops*, pages 879–886. Springer, 2016.

[4] D. Charlet, C. Fredouille, G. Damnati, and G. Senay. Improving speaker identification in tv-shows using person name detection in overlaid text and speech. In *Interspeech*, pages 2778–2782, 2013.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference*, 2014.

[6] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[7] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *Proceedings of the International Conference on Computer Vision*, pages 1559–1566, 2011.

[8] T. Cour, B. Sapp, and B. Taskar. Learning from ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[9] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking pictures: Temporal grouping and dialog-supervised person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 2011.

[11] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proceedings of the 17th British Machine Vision Conference, Edinburgh*, 2006.

[12] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The repere corpus: a multimodal corpus for person recognition. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1102–1107, 2012.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] H. Jiang and E. Learned-Miller. Face detection with the faster R-CNN. *CoRR*, abs/1606.03473, 2016. URL http://arxiv.org/abs/1606.03473.

[15] M. Köstinger, P. Wohlhart, P. Roth, and H. Bischof. Learning to recognize faces from videos and weakly related information cues. In *Advanced Video and Signal based Suerveillance*, 2011.

[16] D. Masip, À. Lapedriza, and J. Vitrià. *Measuring external face appearance for face classification*. INTECH Open Access Publisher, 2007.

[17] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[18] 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.

[19] 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.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, 2015.

[21] J. Poignant, L. Besacier, and G. Quénot. Unsupervised speaker identification in tv broadcast based on written names. *Transactions on Audio, Speech and Language Processing*, 23(1):57–68, 2015.

[22] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *Proceedings of the International Conference on Computer Vision*, 2007.

[23] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people with "their" names using coreference resolution. In *European Conference on Computer Vision (ECCV)*, 2014.

[24] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83, 1995.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[26] M. Rouvier, B. Favre, M. Bendris, D. Charlet, and G. Damnati. Scene understanding for identifying persons in tv shows: beyond face authentication. In *Content-Based Multimedia Indexing, 2014 12th International Workshop on*, pages 1–6. IEEE, 2014.

[27] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[28] M. Tapaswi, M. Baeuml, and R. Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv series. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[29] P. Wohlhart, M. Köstinger, P. M. Roth, and H. Bischof. Multiple instance boosting for face recognition in videos. In *DAGM-Symposium*, 2011.

[30] J. Yang, R. Yan, and A. G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *ACM Multimedia*, 2005.

[31] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015.