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

Arsha Nagrani, Andrew Zisserman
VGG, University of Oxford
Goal

Identify every character in every frame of the video

Useful for:
- Content-based browsing e.g. ‘Fast-forward to when Sherlock first meets John’
- One step closer to story understanding
Previous Approaches

- Rely on transcripts or subtitles as weak supervision
  
  **Transcripts**
  - **who** *speaks* **what**

  **Subtitles**
  - **who** *speaks* **when**
  - **what is spoken** **when**

- This supervision is only weak, hence techniques like Multiple Instance Learning (MIL) are required as well

---

**Everingham et al, 2006**

**Cour et al, 2009**

**Bojanowski et al, 2013**

**Tapaswi et al, 2016**
CNN Face descriptors are marvelous

Can we recognise characters from TV shows using faces of their actors only?

Actor Images from the Web

Raw Facetracks from the video
**Actor Images** are usually taken from red carpet photoshoots
- Frontal
- Good lighting
- Standard expressions

**Character**

Benedict Cumberbatch

Sherlock Holmes
Challenges

Profiles

Extreme poses

Lighting and contrast

Small faces, low resolution

Partial Occlusions
How do we deal with this?

1. Augmentation of actor images
2. Character Context
3. Speech Modality
How do we deal with this?

1. Augmentation of actor images
2. Character Context
3. Speech Modality
1. Augmentation of Actor Images

1. Down-sampling using Bi-cubic interpolation

2. Contrast Adjustment

3. Horizontal Flips
How do we deal with this?

1. Augmentation of actor images
2. Character Context
3. Speech Modality
1. Character Context

- **VGGFace CNN features** are obtained from cropped face regions.
- By retraining on character images we learn hair, make up, expressions of the character.
- Learn the `hairstyle` of the character, not the `hairstyle` of the actor.

Different regions of support
How do we deal with this?

1. Augmentation of actor images
2. Retraining on facetracks from the TV show
3. **Speech Modality**
3. Speech Modality – Voice Classifier

Speaker Identification

Who is the speaker?

Input

Output

SHERLOCK
Speech Modality Pipeline

1. Labels from facetracks

2. Active Speaker Verification

3. Feature Extraction

ASV

4. Classification

SVM classifier

High confidence facetrack

SHERLOCK

1024

Character Face Classifier

VoxCeleb CNN

Spectrogram

VGG,
Dept. of Engineering Science,
University of Oxford

1.

2.

3.

4.
Active Speaker Verification - SyncNet

Voice Feature Extraction - VGGVox

Raw audio signal → 300x512 → 7x7x96 → maxpool → 5x5x256 → maxpool → 3x3x256 → 3x3x256 → 3x3x256 → avgpool → FC7 → FC8

Pretrained on 1,251 speakers (VoxCeleb)

Nagrani, A. Chung, J. S., and Zisserman, A. “VoxCeleb: A large scale speaker identification dataset” INTERSPEECH, 2017
Voice Classifier

- 1 vs rest SVM classifier
- Apply to audio segments where the corresponding face is difficult to identify

Obtain labels for extreme poses and profiles
Putting it all together – Inputs

1. Actor images from the web

   Cast Lists easily available on IMDB

2. Un-labelled facetracks from the TV show

   - Facetracks are obtained using tracking by detection
   - Goal is to update labels using all techniques mentioned so far
Putting it all together

We use three 1-vs-rest SVM face classifiers:

1. **Actor Face Classifier**
   - Trained on augmented actor images only

2. **Character Face Classifier**
   - Trained on character face images, takes into account face context

3. **Character Face Classifier after Voice Correction**
   - Trained on face labels following correction by the voice classifier
Propagation of Confident Labels

1. **Actor Face Classifier**
   - Actor Images
   - Face tracks
   - Most confident labels

2. **Character Face Classifier**
   - Train
   - Correct some labels with Voice Classifier

3. **Character Face Classifier after Voice Correction**
   - Train
Demo of results at each stage
Results at each stage - Sherlock

PR curve, E01

Per Character Accuracy

Very few actor images
Many speaking parts

face(actor) AP: 0.98
face(character) AP: 0.99
face+voice(character) AP: 0.99
Results - Casablanca

Profiles

Dark faces

Partial Occlusions
Comparison to state-of-the-art - Casablanca

- Our method (final): AP: 0.96
- Parkhi ‘15 [2]: AP: 0.93
- Actor face only: AP: 0.89
- Bojanowski ‘13 [1]: AP: 0.75

What have we missed?

- Small and very dark faces
- Extreme occlusion cases where the character is not speaking
- Back of heads
Summary

- Novel approach that eschews transcripts, subtitles or manual annotation
- Multimodal method with both **voice** and **face context** for recognition
- Recognises profiles, partial occlusions and extreme poses
- Beats the state-of-the-art on the Casablanca dataset