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

Arsha NAGRANI, Andrew ZISSERMAN
Visual Geometry Group, Dept. of Engineering Science, University of Oxford

Goal

Character Identification in every frame of the video without a script

- Crucial for story understanding
- Generates metadata for indexing and intelligent fast forwards

Contributions
1. No transcripts or subtitles
2. Novel method using actor images available online
3. Face context to learn character appearance
4. Active Speaker Detection and Speaker Identification to deal with extreme profiles and occlusions

Approach

Actor images are downloaded from the web using the ‘cast list’.

Three Main Stages
1. Actor face classifier
2. Character face classifier
3. Character face + voice classifier

Face Modality

Facetracks are obtained using tracking by detection.

CNN features are obtained from cropped face regions.

Learn the ‘hairstyle’ of the character but not the actor.

Different regions of support

Voice Modality

When is a character speaking?

Correlation between the mouth motion and audio stream is used to perform Active Speaker Verification. This is done using SyncNet.

This gives labels to create a voice model built on audio segments.

Building Voice Models

Datasets

- Ep 1 – 3, TV series
- Crime drama
- Many principal characters
- High resolution, modern

- Full length Feature film
- Black and White
- Many principal & background characters
- Low res, released in 1942

Results

High accuracy on the Sherlock dataset

Per Character Accuracy

Large improvements for speaking characters with few actor images – ‘MIKE’

References: