Quality based Frame Selection for Face Clustering in News Video

Kaneswaran Anantharajah, Simon Denman, Dian Tjondronegoro, Sridha Sridharan, Clinton Fookes and Xufeng Guo
Science and Engineering Faculty,
Queensland University of Technology,
GPO Box 2434, 2 George St., Brisbane, Queensland 4001.
{*k.anantharajah*, *s.denman*, *dian*, *s.sridharan*, *c.fookes*, *felix.guo*}@qut.edu.au

***Abstract*—Clustering identities in a broadcast video is a useful task to aid in video annotation and retrieval. Quality based frame selection is a crucial task in video face clustering, to both improve the clustering performance and reduce the computational cost. We present a frame work that selects the highest quality frames available in a video to cluster the face. This frame selection technique is based on low level and high level features (face symmetry, sharpness, contrast and brightness) to select the highest quality facial images available in a face sequence for clustering. We also consider the temporal distribution of the faces to ensure that selected faces are taken at times distributed throughout the sequence. Normalized feature scores are fused and frames with high quality scores are used in a Local Gabor Binary Pattern Histogram Sequence based face clustering system. We present a news video database to evaluate the clustering system performance. Experiments on the newly created news database show that the proposed method selects the best quality face images in the video sequence, resulting in improved clustering performance.**

I. Introduction

Face clustering in a video is the process of grouping faces that appear in a video based on identity. The identity of people within a video is a key piece of information that can be used to summarize and associate videos, however reliably extracting identity within a single video, and across multiple videos, is difficult due to variations in the environment (i.e. lighting, background, occlusions) and the person themselves (i.e. expression, make up, etc.)

Existing systems tend to rely on heuristics, or simple comparison methods to cluster faces. While significant research has been done in the fields of face recognition [1], [2], face quality assessment [3], [4] and clustering within other domains such as audio (i.e. speech diarisation) [5]; such approaches have not been deployed to cluster faces across a video corpus. Furthermore, existing techniques are typically restricted to clustering within a single video [6], [7], or across multiple videos where subjects faces appear with a near-frontal pose in consistent conditions [8].

In this research, we present an approach to cluster faces across a news video corpus based on selecting high quality faces from long sequences of faces obtained by a face tracking process.

The remainder of the paper is organized as follows. An overview of existing work is presented in Section II; face clustering framework is explained in Section III. In Section IV, we present a new database to facilitate this research, and in Section V, we present the experimental results using this database. We conclude the paper in Section VI.

II. Existing Work

A related task to face clustering is that of speaker diarisation [9], or speech attribution [10], [5]. These systems aim to cluster the speech segments related to a target speaker throughout a single audio file (diarisation) or corpus (attribution). In a speaker diarisation system speech segments corresponding to a speaker are linked without using any prior knowledge. In this work, we seek to develop a similar approach for face.

Various other researchers have proposed face clustering systems [7] for use in video, however they are restricted by assumptions on pose, environment, etc; or they only operate across a single video sequence, rather than a complete corpus. The approach of [11] used k-means to cluster faces within a corpus, however the system required the number of clusters to be defined in advance, and was only evaluated on controlled data.

Pande et al. [6] proposed a method to cluster the faces in a video using a holistic comparison of the face that captured multiple poses, however this approach was limited to clustering within a single video, meaning appearance variations are limited. A similar system was proposed by Elkhoury et al. [12] who use cloth features in addition to facial appearance. However, this approach was also limited by the use of heuristic rules to select a single instance of the face for modeling. Like [6], the system of [12] was only used to cluster faces within a single video.

One possible avenue to improve performance is to use quality measures to select the optimal faces for clustering, and use face recognition to match clusters to one another. Head pose, tilt, brightness, sharpness, resolution, openness of the eye, direction of the eyes and closeness of the mouth features are used to extract high quality face images appearing in a surveillance video [13] and extracted faces are used for verification by a human operator. In order to assess the quality