Robust Automatic Face Clustering in News Video

Kaneswaran Anantharajah, Simon Denman, Dian Tjondronegoro, Sridha Sridharan, Clinton Fookes
Queensland University of Technology (QUT)
k.anantharajah, *s.denman*, *dian*, *s.sridharan*, *c.fookes*@qut.edu.au

***Abstract*—**

Clustering identities in a video is a useful task to aid in video search, annotation and retrieval, and cast identification. However, reliably clustering faces across multiple videos is challenging task due to variations in the appearance of the faces, as videos are captured in an uncontrolled environment. A person's appearance may vary due to session variations including: lighting and background changes, occlusions, changes in expression and make up.

In this paper we propose the novel Local Total Variability Modelling (Local TVM) approach to cluster faces across a news video corpus; and incorporate this into a novel two stage video clustering system. We first cluster faces within a single video using colour, spatial and temporal cues; after which we use face track modelling and hierarchical agglomerative clustering to cluster faces across the entire corpus. We compare different face recognition approaches within this framework. Experiments on a news video database show that the Local TVM technique is able effectively model the session variation observed in the data, resulting in improved clustering performance, with much greater computational efficiency than other methods.

I. Introduction

The identity of people within a video is a key piece of information that can be used to search for, summarize and associate videos. Through knowledge of who is present in a video, various applications including media monitoring to identify the time allocated to a particular person in a news broadcast, or editorial support for journalists to find videos related to a particular person, become possible.

To gather identity information across a video corpus, individuals need to be clustered using a biometric such as face. Reliably clustering faces across multiple videos is a challenging task due to variations in the appearance of the faces, as videos are captured in an uncontrolled environment. A person's appearance may vary between images due to session variations including: lighting changes, background changes, occlusions, changes in expression and make up. Within a single video, cues aside from biometric identity are present that can be used to cluster faces. Within a single news video, footage will often cut between several scenes and people, with the same scene/person revisited several times. Furthermore, the target number of people in either a single video or across a corpus is unknown. Consideration also needs to be given to time constraints, as any approach needs to be able to scale to a large number of videos.

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] and clustering within other domains such as audio (i.e. speech diarisation) [3]; 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 [4, 5], or across multiple videos where subject's faces appear with a near-frontal pose in consistent conditions [6].

In this research, we propose a system to cluster faces in broadcast video using the novel Local Total Variability Modelling (TVM) based face recognition approach. We first cluster faces within a single video by using simple cues such as colour, spatial and temporal information; after which we use novel Local TVM face recognition and hierarchical agglomerative clustering (HAC) to cluster faces across the entire corpus. We compare our proposed Local TVM method with six different face recognition approaches using this framework, and show that the Local TVM face clustering approach offers improved performance over the Local GMM-Free Parts (GMM-FP), Local Intersession Variability Modelling (ISV), Probabilistic Linear Discriminant Analysis (PLDA) and TVM approaches. Further, we shows that the TVM clustering system is also orders of magnitude faster when comparing sequences of faces compared to other approaches.

The remainder of the paper is organized as follows. An overview of existing work is presented in Section II; the face clustering framework is explained in Section III. In Section IV, we present the database and evaluation protocols used in this experiment; and in Section V, we present the experimental results. We conclude the paper in Section VI.

II. Existing Work

Face clustering frameworks consist of three common steps: face detection, feature extraction, and clustering. First, the faces present in a video are detected, after which features are extracted from the faces, and the similarity between pairs of faces (or face tracks) is computed. Finally, a clustering algorithm is used to merge the detected faces. Various researchers have proposed face clustering systems [4, 5, 7–9] for video, however they are either restricted by assumptions on pose, environment, etc; or they only operate across a single video, rather than a complete corpus.

Pande et al. [4] 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, limiting appearance variations. A similar system was proposed by [7] who used cloth features in addition to facial appearance. However, this approach was also limited by the use of heuristic rules to select a single