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

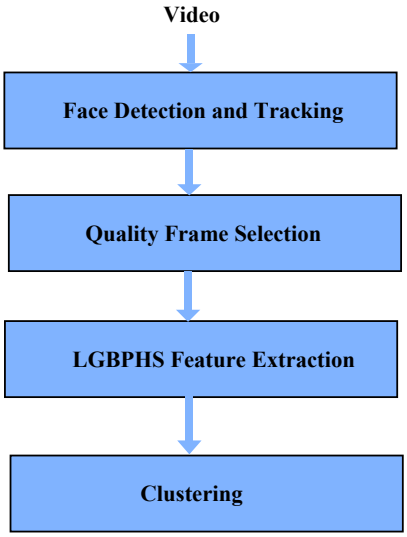

Fig. 1. Face clustering framework.

of a facial image, the pose of the face, lighting, distance from the user to the camera, illumination intensity and sharpness are considered in [14]. In the video based face authentication system of [3], sharpness of the image and face symmetry features are used to select best quality frames.

Barr et al. [8] proposed a framework to detect individuals appearing unusually across multiple videos. The proposed method was tested using videos captured in outdoor and indoor environments, where subjects faces appear with a near-frontal pose. However this method doesn't consider factors important in diarisation such as cluster coverage (i.e. how many of the faces belonging a given subject are captured in the cluster for that subject). Furthermore, evaluation is limited with only 5 of the 90 subjects appearing in multiple videos. In our method we present a face clustering method with quality frame selection and evaluated with 89 subjects with 56 news videos and 50% of subjects are apearing across multiple videos.

III. Face Clustering Framework

The face clustering process employed in our approach is shown in Figure 1. Faces are initially located and tracked in a video sequence, after which they are optionally clustered within the video using cues such as scene changes and local appearance. High quality frames are selected from the resultant set of face tracks, and LGBPHS feature extraction is used to model this faces, and cluster them between videos. Each of these processes is explained in detail in the following subsections.

A. Face Detection and Tracking

In this system frontal faces are detected using a Haar cascade based frontal face detection [15], [16] system. Eye positions of the faces are detected using Haar cascades as well, and detected eye positions are used to normalize the face image to a consistent size (130×150 pixels). Image intensity is normalized using histogram equalization. Faces that appear with a high amount of overlap in successive frames are grouped to form a set of face tracks, which we later seek to merge.

B. Quality Frame Selection

We consider the quality measures based on face symmetry, sharpness, contrast and brightness, as well as a fusion of all four measures. Details on each of this are presented in the following subsections.

a) Face Symmetry: The face asymmetry coefficient proposed in [3] is used as a frontal face and upright face image quality factor. The face asymmetry coefficient, $\alpha(I)$, is calculated as follows,

$$\alpha(I) = \frac{\left\| I - I^f \right\|}{\|I\|}, \tag{1}$$

where, $\left\| I - I^f \right\|$ is the first norm of the difference between image I and I^f is the horizontally flipped version of image I.

Let $\widehat{\alpha_p^{track}}$ be the mean value of the asymmetry coefficient of the normalized face images from the p^{th} face track sequence, then the relative asymmetry coefficient of an image, $\alpha^{relative}$, is calculated as follows,

$$\alpha^{relative}(I) = \alpha(I) - \widehat{\alpha_p^{track}} \tag{2}$$

The asymmetry feature value of the i^{th} image, AF_i, is given by,

$$AF_i = \left[\frac{\alpha(I)}{\sigma_\alpha} + \frac{\alpha^{relative}(I)}{\sigma_\alpha^{relative}} \right], \tag{3}$$

where σ_α and $\sigma_\alpha^{relative}$ are standard deviations of the raw and relative asymmetry measures. Then the normalized symmetry feature, Q_1, is calculated as follows,

$$Q_1 = 1 - \frac{AF_i}{AF_{max}}, \tag{4}$$

where, AF_{max} is the maximum asymmetry feature value of the normalized image in the p^{th} face track. This feature lead to high value for frontal face with no inplane rotation.

b) Sharpness: In blurry images, facial details are not visible, thus the sharpness of the image is considered as a feature in our system. The sharpness of the i^{th} image, S_i, is calculated based on [13] as follows,

$$S_i = \frac{\sum_{x=1}^{x=M} \sum_{y=1}^{y=N} |I - LP(I)|}{M \times N}, \tag{5}$$

where, $LP(I)$ is a Gaussian low-pass filtered image, and M and N, are the height and width of the image respectively.

The normalized sharpness value, Q_2, is calculated as follows,

$$Q_2 = \frac{S_i}{S_{max}}, \tag{6}$$

where, S_{max} is the maximum sharpness value of the normalized image in the i^{th} video sequence.

c) Contrast: The contrast value of the i^{th} image, C_i, is calculated based on [17] as follows,

$$C_i = \frac{B_{q3} - B_{q1}}{I_r}, \tag{7}$$

where, B_{q3} and B_{q1} are histogram bins at which a cumulative histogram have 75% and 25% of the maximum value. I_r is the possible intensity range of the image. Then normalized contrast feature value Q_3 of a face region is calculated as follows,

$$Q_3 = \frac{C_i}{C_{max}}, \tag{8}$$

where, C_{max} is the maximum contrast value of the image in the p^{th} face track.

d) Brightness: The brightness value of the i^{th} face image, B_i, is calculated as follows,

$$B_i = \frac{\sum_{x=1}^{x=M} \sum_{y=1}^{y=N} 0.2989 * R + 0.5866 * G + 0.1145 * B}{M \times N}, \tag{9}$$

where, R, G and B are the red, green and blue components of the image in the RGB colour space. The normalized brightness feature, Q_4, of a normalized face image is calculated as follows,

$$Q_4 = \frac{B_i}{B_{max}}, \tag{10}$$

where, B_{max} is the maximum brightness value of the image in the p^{th} face track.

e) Fusion: We use a weighted summation method to fuse the normalized feature scores, and face images with high score are selected to represent the face track in the clustering process.

f) Temporal Separation: In a video, similar quality face images tend to appear together with small variations. We use a method to ensure that the frames that are selected are not only high quality, but are also taken from different points in the video.

$$\begin{cases} \text{Face images with high score are selected. In this case temporal separation is not considered to avoid low quality frames to represent the face track.} & \text{if } b < \delta \times c \\ \text{Face images with top } \lambda\% \text{ highest. scores are considered in quality frame selection process. These high quality face images are arranged in a temoporal order. Then } i^{th} \text{ face sample, } F_i, \text{ is selected.} & otherwise \end{cases}$$

$$\lambda = \delta \times 100, \tag{11}$$

$$F_i = \frac{\delta \times c \times i}{b} \tag{12}$$

where, b is the number of face samples represent the face track in the clustering process, δ is the quality selection factor and c is number of faces in the p^{th} face track. In this experiment we use $\delta = 0.5$.

C. LGBPHS Feature Extraction

We use the local gabor binary pattern histogram sequence (LGBPHS) [2] to represent the face images in the feature domain. LGBPHS features are extracted from selected face images in a face track. In order to extract LGBPHS features, the normalized face is convolved with five scale and eight orientation Gabor filters, and the resultant gabor magnitude pictures are encoded using local binary patterns. The local gabor binary pattern is divided into 10×10 non overlapping regions and 256 bin histograms are extracted from each region. The histogram intersection value is used to compare two face features. The histogram intersection, $T(p, q)$, of histograms p and q is calculated as follows,

$$T(p, q) = \sum_{k=1}^{k=n} \min(p_k, q_k), \tag{13}$$

where, p, and q are histograms, each containing n bins.

The similarity of two LGBPHS features, $d\left(i, i^i\right)$, is calculated as follows,

$$d\left(i, i^{'}\right) = \sum_{\gamma=0}^{\gamma=7} \sum_{\nu=0}^{\nu=4} \sum_{w=0}^{w=m-1} T\left(i_{\gamma,\nu,w}, i^{'}_{\gamma,\nu,w}\right), \tag{14}$$

where, $i_{\mu,\nu,w}$ and $i^{'}_{\mu,\nu,w}$ are two histograms in the LGBPHS sequence; m is the index of the window; and γ and ν are the Gabor filter orientation and scale respectively.

Sets of faces (i.e. face tracks) are compared using the average similarity of all pair comparisons,

$$d_{av} = \frac{\sum_{f=1}^{f=a} d\left(i_f, i^{'}_f\right)}{a} \tag{15}$$

where,d_{av} is the average distance between two face tracks, and a is the number of faces selected from each face tracks.

D. Clustering faces across multiple videos

The hierarchical agglomerative clustering (HAC) technique [18] has been used in this system as it has shown good performance in speaker diarisation tasks, and is flexible in that the number of clusters is not determined prior to the clustering process. The HAC algorithm works as follows,

1) Initialize all the points as a cluster.
2) Find the nearest cluster pair based on similarity measures and merge. In this experiment we use the complete linkage criteria.
3) Repeat step 2 and terminate the procedure when intergroup similarities exceeds the optimal pre defined threshold.

The similarity between two observations set X and Y is calculated using complete linkage criteria as below,

$$d_{cl} = \max\{d_{av}(x, y) : x \in X, y \in Y\}, \tag{16}$$

where, x and y are face tracks.

We take the face tracks that result from the face tracking process of III-A, and merge these using HAC.

IV. News Video Database

In the existing Honda/UCSD [19] and YouTube celebrities [20] database only one subject is available in a video sequence. However in a broadcast video multiple subjects appear within a single clip, and often simultaneously. Thus, evaluating the proposed clustering system using broadcast video is essential. News videos related to Australian politics have been extracted from Fairfax news videos to form a small video corpus. Ground truth was labeled in order to evaluate face clustering performance. These news videos were recorded in indoor and outdoor environments, and show wide variations in illumination, pose, and clutter. Figure 2 shows frames that contains images with wide variations in illumination. Subjects in these videos also show variations in pose, facial expression, the environment and level of occlusions, as shown in Figures 3, 4, 5 and 6 respectively. In this database the subjects identity and face bounding box location in a frame is labeled. In a video sequence subjects appearing for a short periods of time are ignored and only subjects appearing for a long time are labeled (prominent subjects). Frame numbers in which subject appearance starts and ends, and subject face locations are annotated at every tenth frames (locations are interpolated for intermediate frames). When considering subject appearance, we annotate the facial bounding box for side profile, half profile and profile face images as well (although at present, only frontal faces are detected and tracked by the system).

This database consists of 56 news videos and the total length of video is 119 minutes. This database consists of $167,014$ annotated faces of 89 prominent people. We consider a prominent face to be one that appears for a minimum of 48 frames, and be clearly visible (though not necessarily front on and un-occluded). Examples of prominent and non-prominent faces are shown in Figure 7. In this database 50% of subjects appear across multiple videos including two subjects who appear in 27 videos.

V. Experimental Results

A. Clustering Performance Metrics

Cluster purity and cluster coverage [5] evaluation metrics are used to evaluate the face clustering performance. To obtain these measures, each cluster is analyzed and labeled with its most frequent face image identity. Purity of a cluster, P, is calculated as follows,

$$P = \frac{N^i}{N_t^i}, \tag{17}$$

where, N^i is the number of labeled faces available in the i^{th} cluster and N_t^i is the total number of face images available in the i^{th} cluster.

For each person j, the cluster containing the highest number of the j^{th} person's faces, $\max(N_j)$ is calculated. Then cluster coverage, C, is calculated as follows,

$$C = \frac{\max(N_j)}{N_t^j}, \tag{18}$$

where, N_t^j is the j^{th} person's total number of faces available in the video corpus according to the manual annotation.

Average purity, P_w, and average coverage, C_w, values are used to evaluate the face clustering system performance and are calculated as follows,

$$P_w = \frac{\sum_{t=1}^{t=F} N_t \times P_t}{\sum_{t=1}^{t=F} N_t}, \tag{19}$$

where, P_t is the t^{th} cluster formed by faces detected from video corpus and N_t is the total number of faces available in the cluster t; and F is the total number of face clusters using a detected faces in the video corpus; and,

$$C_w = \frac{\sum_{s=1}^{s=M} R_s \times C_s}{\sum_{s=1}^{s=M} R_s}, \tag{20}$$

where, C_s is the s^{th} subject's coverage from the video corpus and R_s is the total number of s^{th} subject's faces available in the video corpus according to the ground truth; and M is the total number of subjects appearing in the video corpus.

B. Experiment Results

The face clustering system performance is evaluated across a large corpus of videos. Diarisation performance within a single video is shown in Table I. We observe that very high purity is achieved while only a moderate level of coverage is attained. This is to be expected as there is no clustering actually performed within the video. Rather, the face tracking simply outputs a set of face sequences. As such, multiple instances of the same person are not grouped, and the vast majority of faces tracks only contain a single identity. Importantly, the high purity means that clusters being used in the within video clustering system predominately consist of only a single identity, which will aid the within video clustering.

We evaluate the performance of a face clustering system across a video corpus using each quality measure individually as well as the fused combination, and a selection of faces based on simply selecting an equally spaced set from the face track. In our experiment we have chosen 5 and 10 faces to represent face track in the clustering process. Figures 8 and 9 show the face clustering performance across a news video corpus for sets of 5 faces and Figures 10 and 11 show the face clustering performance across a news video corpus for sets of 10 faces. These figures show the trade off between purity and coverage as the final merging threshold is varied. Figure 12 compares the methods that yields best clustering performance for sets of 5 and 10 faces.

Clustering system performance at three operating points (a particular threshold value within the HAC algorithm) are shown in Tables II, III and IV, for all evaluated systems. We evaluate the system at the threshold that yields the correct (or closest to the correct) number of clusters (see Table II), at an operating point of P_w=90% (see Table III) and at the threshold that yields $C_w = P_w$ (see Table IV). We argue that for a diarisation system, the cost of an incorrect merge is greater than the cost of a miss, as it is easier for a human operator

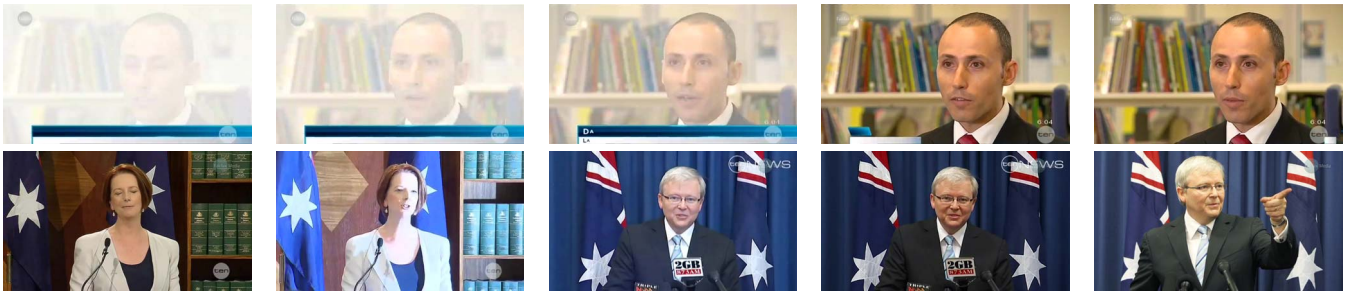

Fig. 2. frames that contains images with wide illumination variation.

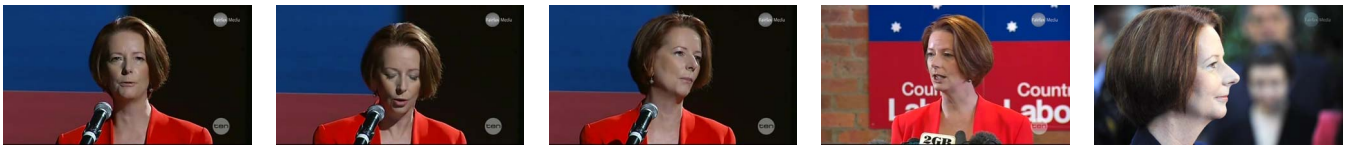

Fig. 3. frames that contains images pose variation.

Fig. 4. Different Facial Expression.

Fig. 5. Frames Captured in Indoor and Outdoor Environment and contains Multiple Faces.

Fig. 6. Partially Occluded Faces.

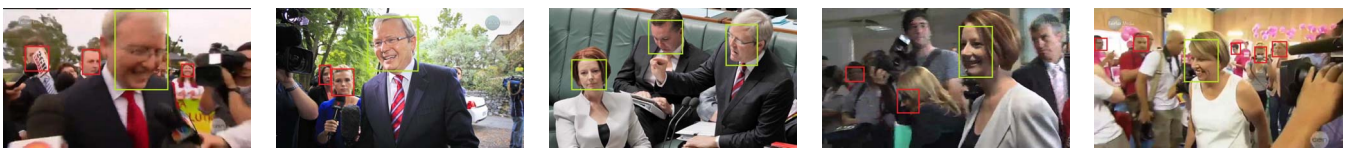

Fig. 7. Prominent faces are marked with green box and non prominent faces are marked with red box.

	Clustering Performance Coverage	Purity
Within video clustering performance	0.39	0.99

TABLE I
WITHIN VIDEO CLUSTERING PERFORMANCE ON NEWS VIDEO DATABASE

to later merge two clusters than separate two potentially very large clusters that have been incorrectly grouped. However, it is also important to consider performance when the correct number of clusters is selected. In total, 776 face cluster are created as a result of the face tracking process, which are the input to the within video clustering.

From Table II, we can observe that the combination of the fused quality metrics and temporal spacing with 5 faces yields the best performance. The use of 5 faces consistently outperforms the system using 10. This can be attributed to both inclusion of low quality faces, and the greater variety of poses within the face sets, which degrades performance.

When we consider an operating point of P_w=90%, the equally spaced 5 face set obtains best performance with 30.6% coverage and 344 clusters. However, the fused quality approaches achieve similar, albeit slightly lower, performace. Comparing 5 and 10 image face sets, performnace varies across the systems. In all cases, the final result is severly underclustered.

When we consider the threshold that yields $C_w = P_w$ (see Table IV), we observe that as with Table II the fusion with temporal spacing approach for face sets of 5 images performs best. However the use of brightness alone performs best with face sets of 10 images. We note that for face set of 5 frames, the system has a tendency to over-cluster (i.e. return fewer clusters than are actually present) at this operating point, while the 10 frame system under-clusters slightly.

Overall, experimental results show that face clustering performance can be improved through the use of quality measures, although, the performance increase is only small. Furthermore, selecting faces that are well distributed in the face track is also important to ensures that variations in the faces are included in the clustering process. However, with the proposed approach we observe that high purity can only be achieved with severe under clustering. Due to the highly varied subject poses within the database, and the fact that the employed face recognition approach does not explicitly consider pose, incorrect merges are easily made resulting in a sharp decrease in purity as coverage increases.

This diarisation system covers 85% of the faces available in the database, because faces appears in half profile and profile poses and are not detected by the frontal face detection. The detection and inclusion of non-prominent faces, as well as false face detections, leads to a maximum purity of 95.8%.

VI. CONCLUSION

In this paper we have investigated the use of quality metrics for face clustering. We have shown that by selecting the highest quality faces from a face track with temporal separation, we

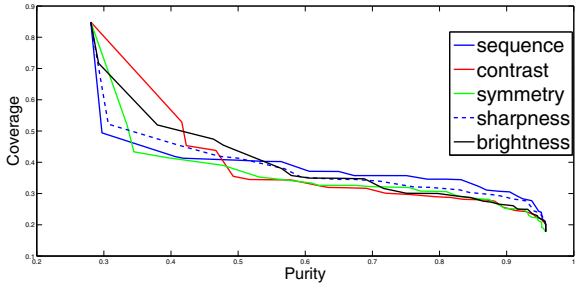

Fig. 8. Coverage Vs purity at different number of clusters when 5 faces are used in clustering

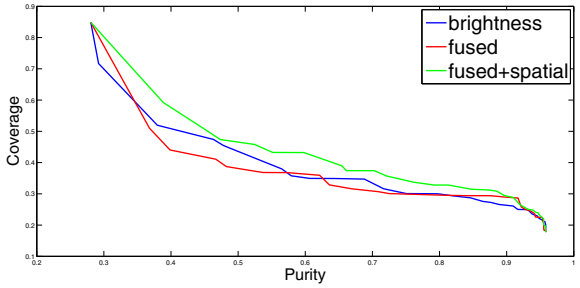

Fig. 9. Coverage Vs purity at different number of clusters when 5 faces are used in clustering

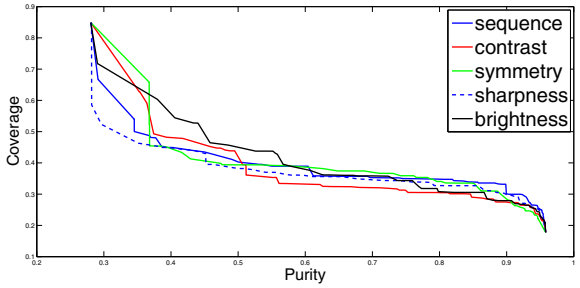

Fig. 10. Coverage Vs purity at different number of clusters when 10 faces are used in clustering

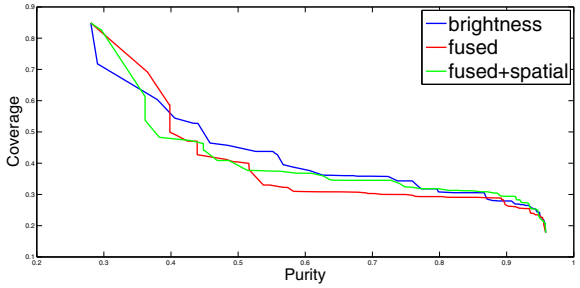

Fig. 11. Coverage Vs purity at different number of clusters when 10 faces are used in clustering

Face Selection Method	Number of Faces = 5			Number of Faces = 10		
	Coverage	Purity	Total	Coverage	Purity	Total
Equally spaced set	40.6	50.6	91.2	48.1	37.0	85.1
Symmetry	38.4	48.7	87.1	45.4	38.5	83.9
Brightness	43.5	50.2	93.7	**53.3**	**42.5**	**95.8**
Sharpness	41.0	50.6	91.6	44.7	40.0	84.7
Contrast	35.5	49.3	84.8	48.0	40.5	88.5
Fusing all features	38.3	49.6	87.9	48.1	41.5	89.6
Fusing + temporal spacing	**44.8**	**53.5**	**98.3**	47.6	41.5	89.1

TABLE II
CLUSTERING SYSTEM PERFORMANCE WHEN CLUSTER SIZE = 89

Face Selection Method	Number of F aces = 5		Number of Faces = 10	
	Coverage when Purity = 90%	Number of Clusters	Coverage when Purity = 90%	Number of Clusters
Equally spaced set	**30.6**	344	**30.0**	338
Symmetry	25.2	415	28.4	353
Brightness	26.4	343	27.9	352
Sharpness	28.8	340	29.8	329
Contrast	25.3	476	27.5	367
Fusing all features	29.3	362	26.5	371
Fusing + temporal spacing	29.0	359	29.4	330

TABLE III
CLUSTERING SYSTEM PERFORMANCE WHEN PURITY = 90%

Face Selection Method	Number of Faces = 5		Number of Faces = 10	
	Coverage = Purity	Number of Clusters	Coverage = Purity	Number of Clusters
Equally spaced set	41.5	81	43.8	96
Symmetry	41.0	82	42.3	92
Brightness	46.8	79	**46.3**	95
Sharpness	43.8	80	43.8	91
Contrast	44.6	80	45.5	92
Fusing all features	42.8	79	43.9	93
Fusing + temporal spacing	**47.3**	81	44.8	91

TABLE IV
CLUSTERING SYSTEM PERFORMANCE WHEN PURITY = COVERAGE

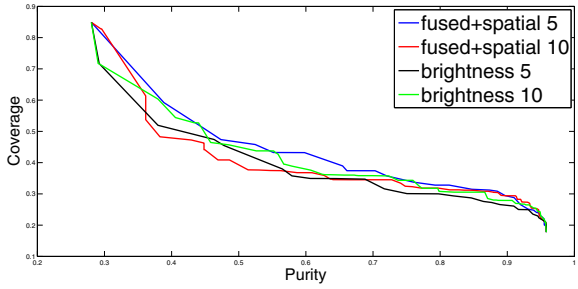

Fig. 12. Best clustering performance for sets of 5 and 10 faces at different number of clusters

can achieve better clustering coverage while maintaining high purity. To cope with the wide variation present in broadcast video, alternate face recognition approaches that incorporate session variability modeling [21] will be investigated for use in face clustering. Approaches to cluster faces within a video using other cues such as scene changes and other video cues will also be investigated.

ACKNOWLEDGMENT

This paper was based on research conducted through the Australian Research Council (ARC) Linkage Grant No: LP0991238 and the follow-up applied research based on Australian Broadcast data conducted through the Cooperative Research Centre for Smart Services.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[2] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, oct. 2005, pp. 786 – 791 Vol. 1.

[3] E. Argones RÃºa, J. Alba Castro, and C. GarcÃa Mateo, "Quality-based score normalization and frame selection for video-based person authentication," in *Biometrics and Identity Management*, ser. Lecture Notes in Computer Science, B. Schouten, N. Juul, A. Drygajlo, and M. Tistarelli, Eds. Springer Berlin / Heidelberg, 2008, vol. 5372, pp. 1–9.

[4] K. Anantharajah, S. Denman, S. Sridharan, C. Fookes, and D. Tjondronegoro, "Quality based frame selection for video face recognition," in *Signal Processing and Communication Systems (ICSPCS), 2012 6th International Conference on*, 2012, pp. 1–5.

[5] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4185–4188.

[6] N. Pande, M. Jain, D. Kapil, and P. Guha, "The video face book," in *Proceedings of the 18th international conference on Advances in Multimedia Modeling*, ser. MMM'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 495–506. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-27355-1_46

[7] S. Foucher and L. Gagnon, "Automatic detection and clustering of actor faces based on spectral clustering techniques," in *Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on*, may 2007, pp. 113 –122.

[8] J. Barr, K. Bowyer, and P. Flynn, "Detecting questionable observers using face track clustering," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, jan. 2011, pp. 182 –189.

[9] V. B. Le, O. Mella, D. Fohr *et al.*, "Speaker diarization using normalized cross likelihood ratio."

[10] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech 2011*, Florence, Italy, August 2011. [Online]. Available: http://eprints.qut.edu.au/43351/

[11] P. Huang, Y. Wang, and M. Shao, "A new method for multi-view face clustering in video sequence," in *Data Mining Workshops, 2008. ICDMW '08. IEEE International Conference on*, dec. 2008, pp. 869 –873.

[12] E. El Khoury, C. Senac, and P. Joly, "Face-and-clothing based people clustering in video content," in *Proceedings of the international conference on Multimedia information retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 295–304. [Online]. Available: http://doi.acm.org/10.1145/1743384.1743435

[13] K. Nasrollahi, T. B. Moeslund, and M. Rahmati, "Summarization of surveillance video sequences using face quality assessment, 11(2), p.207." *International journal of image and graphics*, vol. 11, pp. 207–233, 2011.

[14] X. Gao, S. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science, S.-W. Lee and S. Li, Eds. Springer Berlin / Heidelberg, 2007, vol. 4642, pp. 242–251.

[15] R. Lienhart, L. Liang, and E. Kuranov, "A detector tree of boosted classifiers for real-time object detection and tracking," in *IEEE ICME2003*, 2003, pp. 277–280.

[16] M. C. Santana, O. Déniz-Suárez, L. Antón-Canalís, and J. Lorenzo-Navarro, "Face and facial feature detection evaluation - performance evaluation of public domain haar detectors for face and facial feature detection." in *VISAPP (2)*, 2008, pp. 167–172.

[17] A. Tripathi, S. Mukhopadhyay, and A. Dhara, "Performance metrics for image contrast," in *Image Information Processing (ICIIP), 2011 International Conference on*, nov. 2011, pp. 1 –4.

[18] J. See and C. Eswaran, "Exemplar extraction using spatio-temporal hierarchical agglomerative clustering for face recognition in video," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. 2011, pp. 1481 –1486.

[19] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303 – 331, 2005.

[20] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[21] C. McCool, R. Wallace, M. McLaren, L. E. Shafey, and S. Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, pp. 117–129(12), September 2013. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2012.0059