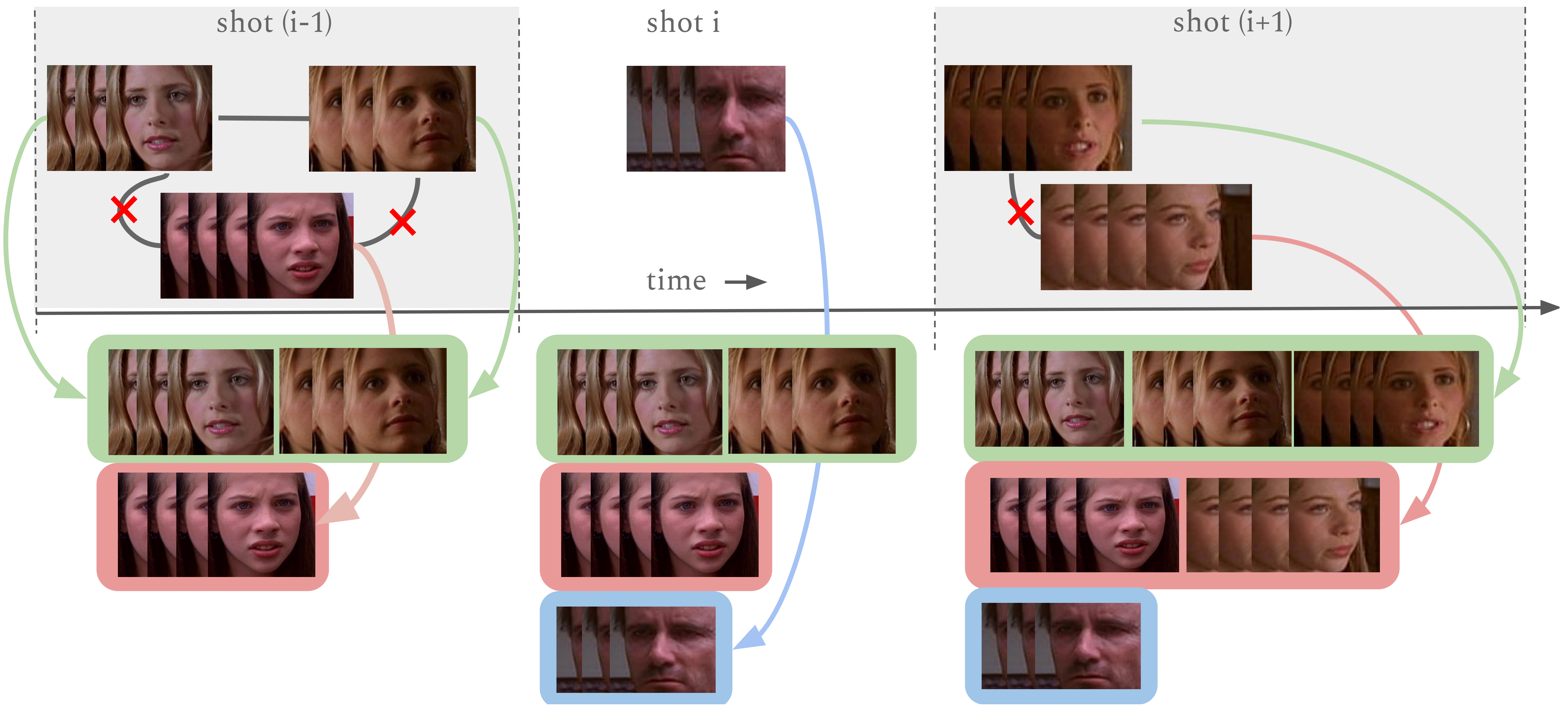
An Online Algorithm for Constrained Face Clustering in Videos

Prakhar Kulshreshtha[1]   Tanaya Guha[1,2]

[1]Indian Institute of Technology, Kanpur   [2]University of Warwick, UK



Motivation & Objective

- Face clustering is challenging: faces exhibit wide variability in scale, pose, illumination, expressions.
- Existing face clustering algorithms are mostly offline - does not help if the complete data is not available or is too large.
- **Accurate *online* face clustering in long, real world videos.**
- Applications: video summarization, indexing, retrieval.

Methodology

- **Shot boundary detection**: Distances between three consecutive frames are used to detect shot boundaries. All frames within a shot are processed together.
- **Feature extraction**: Dlib Face Detector is run on each frame followed by OpenFace to extract FACENET embeddings.
- **Facetrack creation**: Faces in consecutive frames with short spatial distance and feature distance are clubbed together to form facetracks: $\mathcal{V}_k$.
- **Online Clustering**: With K facetracks $\{\mathcal{V}_k\}_{k=1}^{K}$ and set of cluster centers $\mathcal{C}$, construct constraint matrix $\mathbf{Q}$

$$\mathbf{Q}(p,q) = \begin{cases} 0 & \text{if } \mathcal{V}_p \text{ and } \mathcal{V}_q \text{ overlap in time} \\ 1 & \text{otherwise} \end{cases}$$

Construct distance matrix $\mathbf{D}$ and run Algorithm 1

$$\mathbf{D}(l,k) = d(\mathbf{c}_l, \mathcal{V}_k) = 4 - \frac{1}{N_k}\sum_{j=1}^{N_k} \|\mathbf{v}_k^j - \mathbf{c}_l\|_2^2$$

Online Clustering Algorithm

Algorithm 1: Facetrack clustering for a given shot.

```
Input: Face track features in the current shot: $\{\mathcal{V}_k\}_{k=1}^{K}$,
       Initial clusters: $\mathcal{C}$
Output: Updated $\mathcal{C}$
Initialize: $\mathbf{ind} = [1, 2, \ldots, K]$, $\mathbf{W}$ = all-ones matrix.

Compute $\mathbf{Q}$, $\mathbf{D}$ using (1) and (1)
while length($\mathbf{ind}$) $> 0$ do
  if $\mathcal{C}$ not empty && $\max_{l,k}(\mathbf{D} \odot \mathbf{W}) \geq \tau$ then
    $(\hat{l}, \hat{k}) \leftarrow \text{argmax}_{l,k}(\mathbf{D} \odot \mathbf{W})$
    $k^* \leftarrow \mathbf{ind}[\hat{k}]$
    Update cluster center $\mathbf{c}_{\hat{j}}$ with $\mathcal{V}_{k^*}$
  else
    Add new cluster $(\hat{l}, \hat{k}) \leftarrow (L+1, 1)$
    $k^* \leftarrow \mathbf{ind}[\hat{k}]$
    $\mathbf{c}_{new} \leftarrow$ mean($\mathcal{V}_{k^*}$)
    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathbf{c}_{new}$
  end
  Recompute $\mathbf{D}$ for $\mathbf{c}_{\hat{l}}$
  $\mathbf{W}(\hat{l}, :) \leftarrow \mathbf{Q}(\hat{k}, :)$
  Delete $\mathbf{D}[:, \hat{k}], \mathbf{W}[:, \hat{k}], \mathbf{Q}[\hat{k}, :], \mathbf{Q}[:, \hat{k}], \mathbf{ind}[\hat{k}]$
end
```

Performance Evaluation

- *Buffy* database (BF2006) [1]: 229 facetracks of 6 characters (17,337 faces)
- *Notting Hill* database (NH2016) [2]: 277 facetracks of 7 characters (19,278 faces)

Table 1: Comparison with the online face clustering method

	BF2006		**NH2016**	
	TCCRP [3]	Proposed	TCCRP [3]	Proposed
Homogeneity	**0.93**	0.68	**0.92**	0.88
Completeness	0.44	**0.69**	0.44	**0.89**
V measure	0.60	**0.68**	0.58	**0.89**
clusters	57	7	61	7

Table 2: Comparison with the state-of-the-art (offline) clustering methods in terms of clustering accuracy (%)

Method	**BF2006**	**NH2016**
ULDML [4]	49.29	43.82
cHMRF [2]	61.87	47.94
FaceNet + aCNN [5]	**89.90**	90.17
FaceNet + GMM	84.21	73.46
FaceNet + Kmeans	82.92	71.66
Proposed	82.12	**93.84**
Proposed + GMM	93.79	94.17

Summary

- Proposed an online clustering algorithm that performs as good or better than existing online or offline methods.
- Used FACENET embedding for robust representation of faces, and several spatio-temporal constraints to cluster the faces as they appear.
- Achieved high clustering accuracy on two real world video databases.
- Can be extended by allowing online splitting and fusing of clusters.

Qualitative Results

6 character clusters, 1 noisy cluster (7th row) in BF2006.

6 character clusters, 1 noisy cluster (7th row) in NH2016.

Code and Experiments

- Code and experiments available at *https://github.com/ankuPRK/COFC*
- For queries and suggestions please email to *ankuprk@gmail.com*

References

[1] M. Everingham, J. Sivic, and A. Zisserman, “Hello! my name is... buffy–automatic naming of characters in tv video,” in *BMVC*, 2006.

[2] B. Wu, B. Hu, and Q. Ji, “A coupled hidden markov random field model for simultaneous face clustering and tracking in videos,” *Pattern Recognition*, vol. 64, pp. 361–373, 2017.

[3] A. Mitra, S. Biswas, and C. Bhattacharyya, “Bayesian modeling of temporal coherence in videos for entity discovery and summarization,” *IEEE Trans PAMI*, vol. 39, no. 3, pp. 430–443, 2017.

[4] R. G. Cinbis, J. Verbeek, and C. Schmid, “Unsupervised metric learning for face identification in tv video,” in *ICCV*, pp. 1559–1566, 2011.

[5] V. Sharma, M. S. Sarfraz, and R. Stiefelhagen, “A simple and effective technique for face clustering in tv series,”