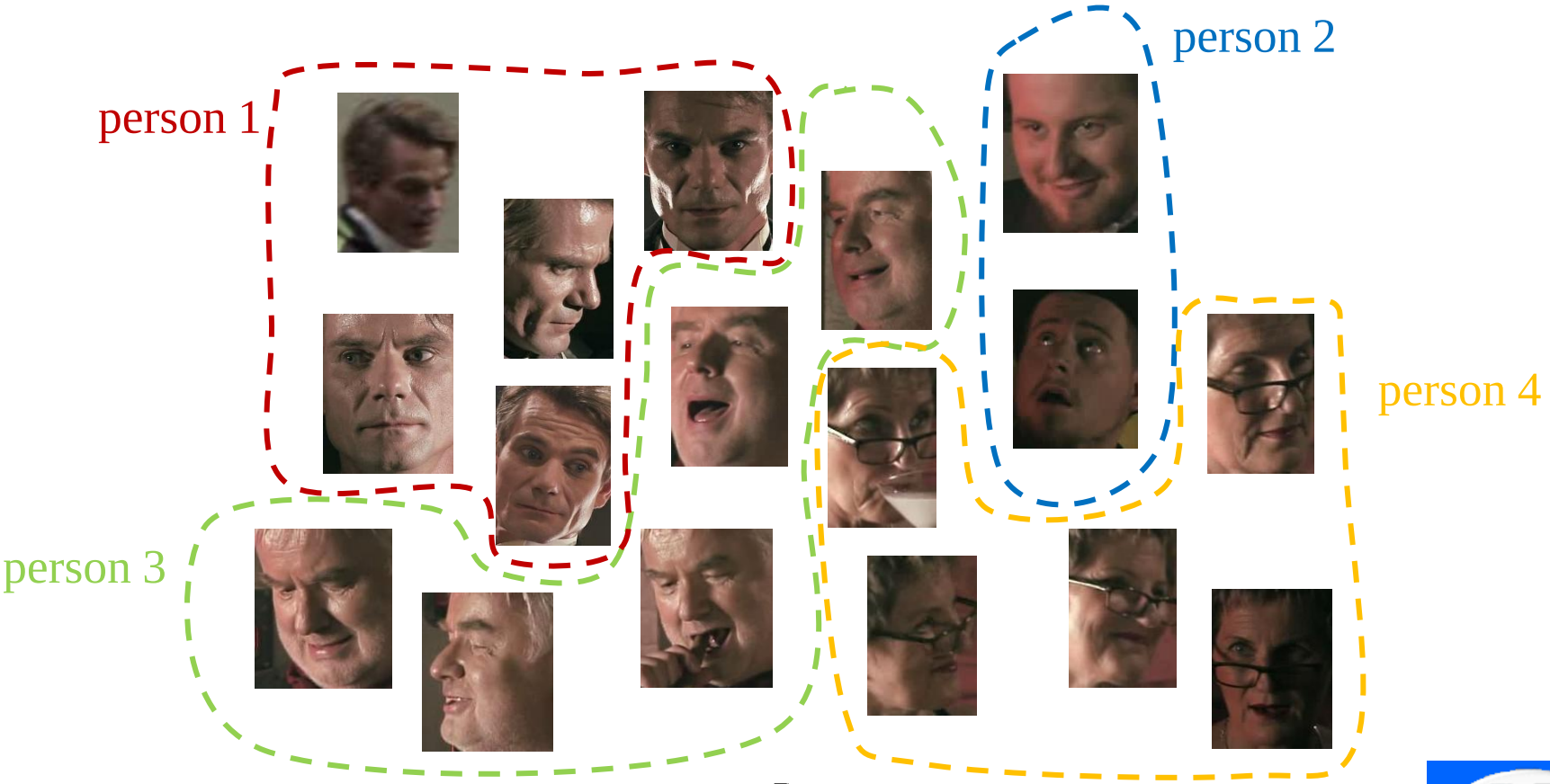
Exploiting Clustering and Disparity Information in Label Propagation on Facial Images

Olga Zoidi
Nikos Nikolaidis
Ioannis Pitas*
AIIA Lab
Aristotle University of Thessaloniki
Greece

IEEE SSCI
16-19/4/2013

Introduction

- Which images belong to the same person?



Introduction

- The successful management of the large amount of information in video archives requires the development of efficient ways for describing and searching the stored content.
- If the user searches for the appearances of a specific actor within a movie, the video annotation should contain the identity labels of actors that appear in each frame.
- An average movie consists of more than 100,000 frames, therefore manual annotation of an entire movie is labor-intensive and time-consuming.
- This problem can be overcome with semiautomatic annotation techniques based on label propagation.

Label Propagation

- $\mathcal{X}_L = \{\mathbf{x_i}\}_{i=1}^{n_l}$: the set of labeled data
- $\mathcal{L} = \{l_j\}_{j=1}^{L}$: the set of labels
- $\mathcal{X}_U = \{\mathbf{x_i}\}_{i=1}^{n_u}$: the set of unlabeled data
- $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_l}, \mathbf{x}_{n_l+1}, \ldots, \mathbf{x}_N\}$: the set of labeled and unlabeled data, where $N = n_l + n_u$
- $\mathbf{Y} = [y_1, \ldots, y_{n_l}, 0, \ldots, 0]^T \in \mathcal{L}^N$: contains the labels of the labeled data in the first n_l positions and takes the value 0 in the last n_u positions.

Label Propagation

- The objective of label propagation is to spread the labels in $\mathcal{L}$ from the set of labeled data $\mathcal{X}_L$ to the set of unlabeled data $\mathcal{X}_U$
- Label propagation methods satisfy the following requirements:
  - they should retain the labels of the initial labeled samples
  - they should assign the same label to similar samples, or to samples that lie to the same structure of the feature space.

Label Propagation

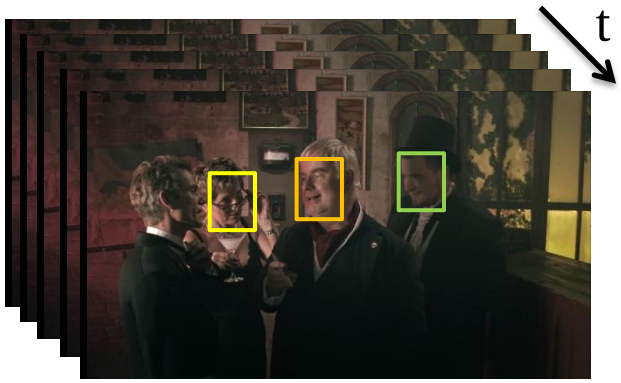
- Label propagation results depend highly on
  - The data representation method (graph construction)
  - The selection of the samples from which label inference should begin

Label Propagation

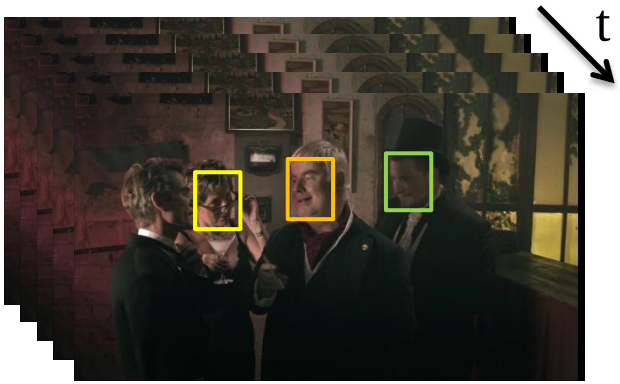
- The proposed method performs label propagation on facial images
- It exploits prior information for the data structure, obtained from the application of a clustering algorithm, for the selection of the facial images from which label inference should begin.
- A sparse graph is constructed according to the Linear Neighborhood Propagation (LNP) method
- Label inference is performed according to an iterative update rule.
- In the case of stereoscopic videos, the classification decision is determined by the combined information of the left and right channels.

Dataset initialization

- Perform image acquisition through automatic face detection and tracking in a video
- For stereoscopic videos, automatic face detection and tracking is performed in the stereo pairs of facial images in the left and right video channels
  - This results to two sets of facial images, for the left and the right video channel



Left channel



Right channel

Dataset initialization

- Restrictions:
  - the facial images that belong to the same trajectory belong to the same person and, therefore, should be assigned the same identity label
    - only the first facial image of each trajectory is placed in the data set
    - The labels are propagated to the remaining facial images of the trajectories.
    - The computational cost is reduced by two orders of magnitude
  - in stereoscopic videos, the facial images in the left and right video channels that depict the same person should be assigned the same label

Labeled dataset initialization

- Application of a clustering algorithm to the data set $\mathcal{X}$
- Label (manually) the median of each cluster with the corresponding person identity
  - Process clusters in decreasing cardinality (largest cardinality clusters first)
  - Only one facial image for each person is labeled
- If a facial image of the same person already exists in $\mathcal{X}_L$ ignore it and continue to the next cluster
- until all clusters are processed or the image of the last person enters the labeled set $\mathcal{X}_L$
- Random selection of the facial images of the remaining persons

Linear Neighborhood Propagation

- Graph construction:
  - each graph node $\mathbf{x}_i$ is reconstructed from its *k*-nearest neighbors with respect to mutual information

$$\mathbf{x}_i = \sum_{i_j : \mathbf{x}_{i_j} \in \mathcal{N}(\mathbf{x}_i)} W_{ii_j} \mathbf{x}_{i_j}$$

$$\text{s.t.} \sum_{i_j : \mathbf{x}_{i_j} \in \mathcal{N}(\mathbf{x}_i)} W_{ii_j} = 1, \quad W_{ii_j} \geq 0$$

  where $\mathcal{N}(\mathbf{x}_i)$ is the neighborhood of node $\mathbf{x}_i$.
  - The weights W_{ij} on the edges of the constructed graph are selected such that they minimize the reconstruction error.

Linear Neighborhood Propagation

- Label inference:
  - Each node incorporates label information both from the neighboring facial images and the assigned label information of its initial state (if any)
  - The matrix **F**:

$$\mathbf{F} = [\mathbf{f}^1, \ldots, \mathbf{f}^L] \in \Re^{N \times L}$$

N: number of images

L: number of labels

assigns in each node one value (score) for each label according to: $\mathbf{F}' = (1-\alpha)(\mathbf{I} - \alpha\mathbf{W})^{-1}\mathbf{Y}$

  - The matrix **Y** contains the labels of the labeled nodes. If the node is unlabeled, the value of **Y** is set 0.

a: the fraction of label Information the node receives from its neighbors

Linear Neighborhood Propagation

- Label inference:
  - The facial image $\mathbf{x}_i$ is assigned a person identification label y_i according to:

$$y_i = \text{argmax}_{l \in \{1,\ldots,L\}} [f_i^1 \quad f_i^2 \quad \ldots \quad f_i^L]$$

Exploiting Stereo Information

- Exploiting stereo information in two ways
  - **Early fusion:** Combine the weight matrices to a single weight matrix: $\mathbf{W}_S = \frac{1}{2}\mathbf{W}_{knnL} + \frac{1}{2}\mathbf{W}_{knnR}$ and perform label propagation to the resulting matrix
  - **Late fusion:** Label inference is performed independently on the left and right video channel and the resulting matrices $\mathbf{F}^L$, $\mathbf{F}^R$ are merged according to:

$$\mathbf{F}_{il}^{max} = \max(\mathbf{F}_{il}^{L}, \mathbf{F}_{il}^{R})$$

  - The facial image is assigned the label according to:

$$y_i = \text{argmax}_{l \in \{1,\ldots,L\}} [f_{i1}^{max} \quad f_{i2}^{max} \quad \ldots \quad f_{iL}^{max}]$$

Experimental results

- Compare the cluster-based classification results to the average algorithm results after 20 runs when the labeled data set is selected randomly.
- Classification accuracy is measured by the *F*-measure

$$F = \sum_{i=1}^{N} \frac{N_i}{N} F_i \qquad F_i = 2\frac{precision_i \cdot recall_i}{precision_i + recall_i}$$

$$precision_i = \frac{|correctly\ classified\ images\ of\ class\ i|}{|classified\ images\ of\ class\ i|}$$

$$recall_i = \frac{|correctly\ classified\ images\ of\ class\ i|}{|images\ of\ class\ i|}$$

3DTVS

AIA Laboratory

Experimental results

- Monocular videos

movie	random	cluster-based
American Beauty	0.9838	1.0000
As Good As It Gets	0.4472	0.5559
Being John Malkovitch	0.9987	1.0000
Big Lebowski	0.9653	1.0000
The Butterfly Effect	0.8579	0.9153
Erin Brockovitch	0.3400	0.4797
Forest Gump	0.9002	1.0000
The Graduate	0.61114	0.6599
I Am Sam	0.9985	1.0000
Indiana Jones and the last crusade	0.9106	0.9838
Kids	0.8939	0.9311
LOR	0.9654	0.9682

  - In all videos the classification accuracy improves when we exploit clustering information

Experimental results

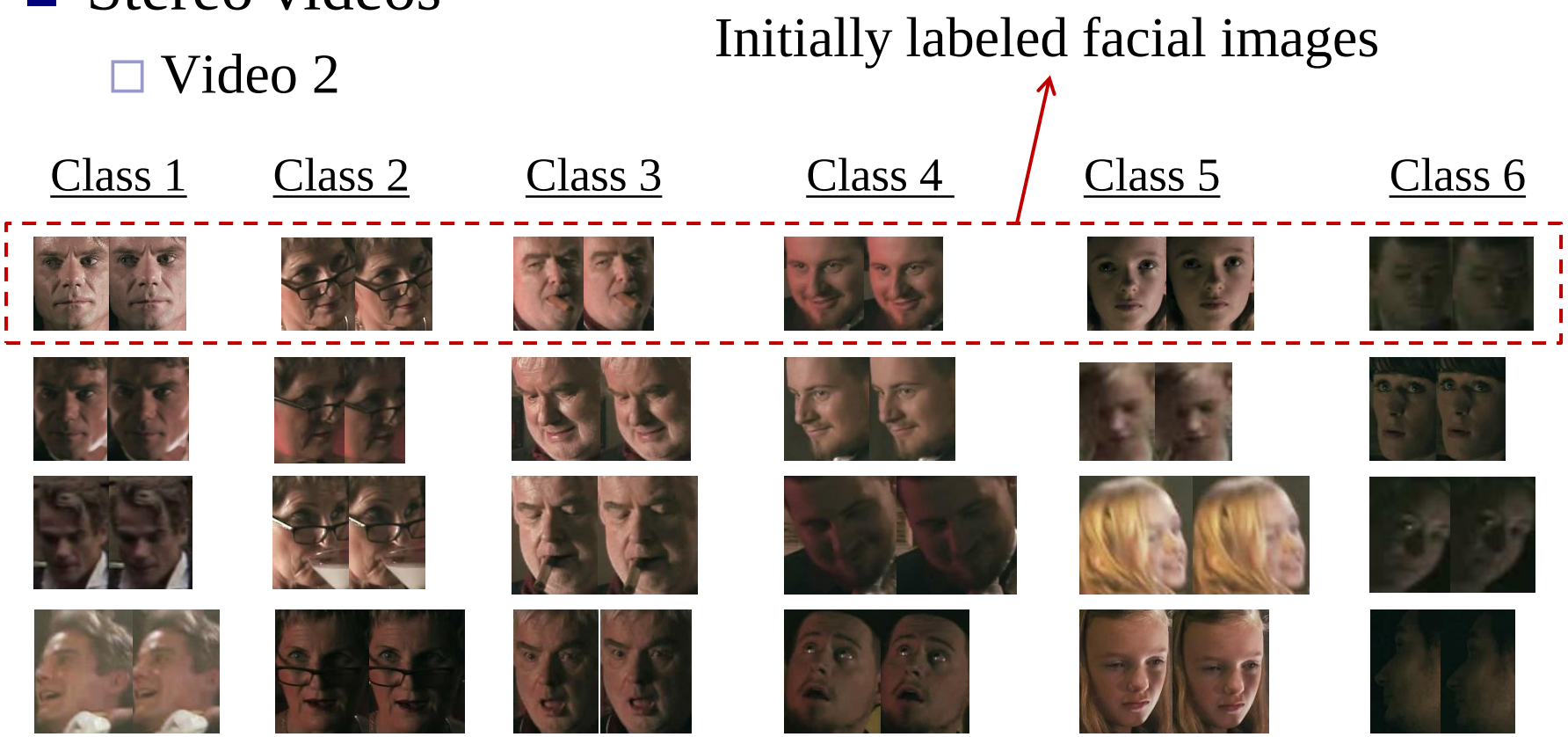
- Stereo videos
  - Video 1: 45 stereo trajectories, 3,805 stereo facial images belonging to 3 individuals.

Experimental results

- Stereo videos
  - Video 2: 195 stereo trajectories, 15,992 stereo facial images belonging to 13 individuals.
    - 8 out of the 13 individuals had few appearances in the video, therefore they were considered as belonging to the same class with the label ’supporting actor’.

Experimental results

- Stereo videos
  - Video 2



Experimental results

- Random initialization

	Single channel		Stereo	
video	F_L	F_R	F_{max}	F_S
video 1	**0.8008**	0.7714	0.7965	0.7963
video 2	0.4412	0.4447	0.4728	**0.4868**

F_L: left channel
F_R: right channel
F_{max}: late fusion
F_S: early fusion

- Cluster-based initialization

	Single channel		Stereo			
video	F_L	F_R	$F_{max}(L)$	$F_{max}(R)$	$F_S(L)$	$F_S(R)$
video 1	**0.8766**	0.8523	**0.8766**	**0.8766**	**0.8766**	**0.8766**
video 2	0.5920	0.5961	0.6210	0.6442	0.6527	**0.6952**

indices L and R indicate the channel whose clustering results were taken into consideration for the initialization of $\mathcal{X}_L$

3DTVS

Experimental results

- The use of stereo information in LNP has a positive influence on the classification accuracy with respect to the single channel LNP, as it enhances it (video2) or makes it more robust (video1).
- The initialization of $\mathcal{X}_L$ according to the clusters' most representative facial images achieves 7-21% better classification accuracy than the random initialization.

Conclusions

- A framework for semi-automatic person identity label propagation on monocular and stereo facial images was presented
- The framework exploits information about the data structure (clustering information) for the initialization of the LNP algorithm.
- Two methods for exploiting stereo information obtained from the left and right channel in the classification decision were introduced based on early and late fusion.
- Experimental results showed the superiority of the proposed cluster-based LNP framework over the state of the art LNP method.

Acknowledgements

- The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS).

http://www.3dtvs-project.eu

3DTVS