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

- \( \mathcal{X}_L = \{x_i\}_{i=1}^{n_l} \): the set of labeled data
- \( \mathcal{L} = \{l_j\}_{j=1}^{L} \): the set of labels
- \( \mathcal{X}_U = \{x_i\}_{i=1}^{n_u} \): the set of unlabeled data
- \( \mathcal{X} = \{x_1, \ldots, x_{n_l}, x_{n_l+1}, \ldots, x_N\} \): 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
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 $x_i$ is reconstructed from its $k$-nearest neighbors with respect to mutual information
    \[
    x_i = \sum_{i,j: x_j \in \mathcal{N}(x_i)} W_{ij} x_j
    \]
  - The weights $W_{ij}$ on the edges of the constructed graph are selected such that they minimize the reconstruction error.

  where $\mathcal{N}(x_i)$ is the neighborhood of node $x_i$. 

  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 $F$:
    \[ F = [f^1, \ldots, f^L] \in \mathbb{R}^{N \times L} \]
    assigns in each node one value (score) for each label according to:
    \[ F' = (1 - \alpha) (I - \alpha W)^{-1} Y \]
  - The matrix $Y$ contains the labels of the labeled nodes. If the node is unlabeled, the value of $Y$ is set 0.

\[ N: \text{number of images} \]
\[ L: \text{number of labels} \]
\[ \alpha: \text{the fraction of label information the node receives from its neighbors} \]
Linear Neighborhood Propagation

- Label inference:
  - The facial image $x_i$ is assigned a person identification label $y_i$ according to:

$$y_i = \arg\max_{l \in \{1, \ldots, L\}} [f_i^1 \quad f_i^2 \quad \ldots \quad f_i^L]$$
Exploiting Stereo Information

Exploiting stereo information in two ways

- **Early fusion:** Combine the weight matrices to a single weight matrix:
  \[ W_S = \frac{1}{2} W_{knnL} + \frac{1}{2} W_{knnR} \]
  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 \( F^L \), \( F^R \) are merged according to:

  \[ F^\text{max}_{il} = \max(F^L_{il}, F^R_{il}) \]

- The facial image is assigned the label according to:

  \[ y_i = \arg\max_{l \in \{1, \ldots, L\}}[f^\text{max}_{i1}, f^\text{max}_{i2}, \ldots, f^\text{max}_{iL}] \]
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

\[
F = \sum_{i=1}^{N} \frac{N_i}{N} F_i \quad F_i = 2 \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i}
\]

\[
\text{precision}_i = \frac{|\text{correctly classified images of class } i|}{|\text{classified images of class } i|}
\]

\[
\text{recall}_i = \frac{|\text{correctly classified images of class } i|}{|\text{images of class } i|}
\]
Experimental results

- Monocular videos

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

- 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.

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="left_images" alt="Images" /></td>
<td><img src="middle_images" alt="Images" /></td>
<td><img src="right_images" alt="Images" /></td>
</tr>
<tr>
<td><img src="left_images" alt="Images" /></td>
<td><img src="middle_images" alt="Images" /></td>
<td><img src="right_images" alt="Images" /></td>
</tr>
</tbody>
</table>
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

Initially labeled facial images

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Class 1 Images" /></td>
<td><img src="image2.png" alt="Class 2 Images" /></td>
<td><img src="image3.png" alt="Class 3 Images" /></td>
<td><img src="image4.png" alt="Class 4 Images" /></td>
<td><img src="image5.png" alt="Class 5 Images" /></td>
<td><img src="image6.png" alt="Class 6 Images" /></td>
</tr>
</tbody>
</table>
Experimental results

- **Random initialization**

<table>
<thead>
<tr>
<th>video</th>
<th>$F_L$</th>
<th>$F_R$</th>
<th>$F_{max}$</th>
<th>$F_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>video 1</td>
<td><strong>0.8008</strong></td>
<td>0.7714</td>
<td>0.7965</td>
<td>0.7963</td>
</tr>
<tr>
<td>video 2</td>
<td>0.4412</td>
<td>0.4447</td>
<td>0.4728</td>
<td><strong>0.4868</strong></td>
</tr>
</tbody>
</table>

- **Cluster-based initialization**

<table>
<thead>
<tr>
<th>video</th>
<th>$F_L$</th>
<th>$F_R$</th>
<th>$F_{max_L}$</th>
<th>$F_{max_R}$</th>
<th>$F_{SL}$</th>
<th>$F_{SR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>video 1</td>
<td><strong>0.8766</strong></td>
<td>0.8523</td>
<td><strong>0.8766</strong></td>
<td><strong>0.8766</strong></td>
<td><strong>0.8766</strong></td>
<td><strong>0.8766</strong></td>
</tr>
<tr>
<td>video 2</td>
<td>0.5920</td>
<td>0.5961</td>
<td>0.6210</td>
<td>0.6442</td>
<td>0.6527</td>
<td><strong>0.6952</strong></td>
</tr>
</tbody>
</table>

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

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