ABSTRACT
Labeling persons appearing in video frames with names detected in a corresponding video transcript helps improving video content annotation and search tasks. We develop a face naming method that learns from labeled and unlabeled examples using iterative label propagation in a graph of connected faces or name-face pairs. The advantage of this method is that it uses few labeled data points and incorporates the unlabeled data points during the learning process. Anchor detection boosts the face naming performance.

1. INTRODUCTION
People play an important role in most video content, and tools for identifying the individuals appearing in a video would be especially helpful. Here, we label news video with textual information extracted from the transcripts. This is a difficult problem, as subtitles usually do not directly describe the video content, but on the contrary are often complementary to the visual information. Moreover, even if the name of a face is explicitly mentioned, the temporal alignment may be relatively weak. In addition, there are often many faces in a frame, and many names in the transcripts of the video, as well as many unnamed faces or names that are mentioned but not displayed. At the same time, the appearance of a person’s face can vary dramatically, due to changes in pose, changes in lighting conditions, facial expressions and partial occlusions. To make the quality of the results acceptable, we propose a semi-supervised scheme, where a few faces of each person in the database are labeled manually, and this information is then propagated to similar unlabeled faces. This is achieved using a graph based algorithm for learning the name-face alignments jointly from labeled and unlabeled examples. By first identifying the news anchor(s), the overall performance can be improved significantly. Variant versions of the label propagation model are proposed that deal with time and space complexity of processing large datasets. In addition, we show the portability of labeled seeds across nine different broadcasts of BBC news.\(^1\)

2. VIDEO PROCESSING
The video frames and related text transcripts are processed with the goal of detecting faces [1], tracking the faces [2] in subsequent frames, detecting person names in the transcripts with a named entity recognizer and tracking the references to a person in the texts with a coreference resolution tool. Anchors may be problematic, since their names are typically mentioned only once, at the very beginning of the news broadcast. We exploit the fact that anchors typically occur over a wide time range and against a typical background (e.g. the news studio). We find the anchor face tracks based on scene similarity using a minimum spanning tree algorithm and assume that repeating scenes represent the studio setting of the anchor. The anchor name is assigned based on the temporal alignment between the name and the anchor face track.

Once the anchor faces are labeled, the remainder of the faces are annotated with a label propagation algorithm. We first manually annotate a small number of name-face seed pairs. Suppose we initially can find name labels for \(l\) faces and the remaining \(u\) faces do not have name labels yet. We denote \((f_1, n_1)\) ... \((f_l, n_l)\) the set of labeled faces where \(N_l = \{n_1,...,n_l\}\), \(n_i \in \{1,...,C\}\) are the name labels. The constraints are the number of distinct names \(C\) is known and all the distinct name labels appear in the set of labeled faces. The set of unlabeled faces can be presented as \((f_{l+1}, n_{l+1})\) ... \((f_{l+u}, n_{l+u})\) where the name labels \(N_u = \{n_{l+1},...,n_{l+u}\}\) are not known yet. Let \(F\) be the set of all faces \(F = \{f_1,...,f_{l+u}\} \in R^D\) described with \(D\) features. We will use \(F\) and \(N_l\) to predict \(N_u\).

We have developed a face similarity metric as an exponential function with a parameter that moderates variations in expressions, occlusions, and unmodeled lighting changes trained on a validation set. If a group of faces are similar, they may have the same name. The level of similarity between these faces determines the confidence that they share the same name. Labeled faces contribute to the naming of

\(^1\)A full version of this paper is available as Phi The Pham, Tinne Tuytelaars and Marie-Francine Moens, Naming People in News Videos with Label Propagation. IEEE MultiMedia 18(3): 44-55 (2011).
an unlabeled face by their names with the confidence estimated by their similarities with this unlabeled face. Moreover, unlabeled faces also affect the labels of each other by their similarity or co-occurrence. To implement the above observation, we build a fully connected graph $G^{l}$ where the nodes are all $l + u$ faces. The weight $w_{ij}$ of the edge between faces $f_i$ and $f_j$ is the similarity between them, and the one-step transition probability $T_{ij}$ from face $j$ to face $i$ can be estimated as the normalized edge weights. All faces have probability distributions over name labels. We define a probability matrix $N^{l}$ of size $(l + u) \times C$ where row $N_{i}^{l}$ represents the probability distribution over all name labels for face $i$. In videos, we notice that faces in the same frame almost always have different names; while faces in the same face track must have the same name. These extra constraints can be exploited in building the connected graph of name-face pairs $G^{nf}$.

After setting up the graph $G^{*}$ (either $G^{l}$ or $G^{nf}$), the transition matrix $T^{*}$ and the label matrix $N^{*}$, we perform the label propagation algorithm as follows:

1. All faces/name-face pairs propagate labels for one step: $N^{*} \leftarrow T^{*} N^{*}$
2. Row normalize $N^{*}$ to maintain the label probability interpretation.
3. Clamp the labeled faces. Repeat from step 2 until convergence of $N^{*}$.

Matrix $N^{*}$ now contains the label distribution for each face or name-face pair. In the experiments below we use for each face the name with highest probability or the alignment of a name-face pair with the labeled seed with highest probability where the probability is above a threshold $\lambda$ (called name assignment threshold) [2]. This refusal to predict strategy leaves some faces unlabeled – in other words these faces refer to the null name.

The size of the graphs $G^{l}$ and $G^{nf}$ grows rapidly with the increase in the number of faces. We have experimented with an approach that only uses the face tracking information, and an approach that implements a sparse matrix. In the first approach, each node in the graph, instead of being a face/name-face pair, is represented by a face track/name-face track pair. In the second approach we restrict a face to be linked with at most $k_{f}$ nearest neighbor faces and a face track to be linked with at most $k_{t}$ nearest neighbor face tracks. This makes $T^{*}$ a real sparse matrix, which we can store efficiently.

3. RESULTS AND DISCUSSION

We perform and compare our experiments on nine BBC news broadcasts recorded from 22-Jun-2008 to 01-Jul-2008. Each broadcast lasts approximately 30 minutes, or 60,000 frames. After the face detection and tracking process, we obtain from each broadcast an average number of 28,000 faces, forming on average 120 face tracks. Three representative faces are selected for each track based on their size (the larger, the better), then by their fitting confidence. After reduction, a broadcast contains on average 496 faces and 116 face tracks. For the name detection and clustering, we obtain from each broadcast an average of 21 unique candidate names. We evaluate here the label propagation in three settings using graph $G^{nf}$: (1) without parameter learning of the face similarity and without anchor detection (denoted as $NF-NP-NA$); (2) with parameter learning but without anchor detection ($NF-WP-NA$); and (3) with both parameter learning and anchor detection ($NF-WP-WA$). The label propagation process is applied on each news broadcast separately. Then, the final face labeling performance is the average of the respective recall and precision values (for the unlabeled faces) on all the broadcasts. To clearly assess the label propagation, we compare the results with the results of a Support Vector Machine (NF-SVM) (optimized with an RBF kernel) and a nearest neighbor classifier (NF-NN) (simulating a k-means clustering where the centroids are based on the labeled seeds) integrating the same constraints as used with graph $G^{nf}$. Table 1 shows that the best algorithm obtains a precision of more than 86% when 70% of the faces are labeled. All results generally outperform the results of the SVM and NN classifiers.

<table>
<thead>
<tr>
<th>Recall</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT-NP-NA</td>
<td>76.15</td>
<td>55.77</td>
<td>50.75</td>
<td>49.78</td>
</tr>
<tr>
<td>NT-WP-NA</td>
<td>68.70</td>
<td>55.48</td>
<td>53.00</td>
<td>49.78</td>
</tr>
<tr>
<td>NT-WP-WA</td>
<td>86.29</td>
<td>86.29</td>
<td>80.89</td>
<td>74.40</td>
</tr>
<tr>
<td>NT-SVM</td>
<td>61.06</td>
<td>57.79</td>
<td>52.00</td>
<td>47.09</td>
</tr>
<tr>
<td>NT-kmeans</td>
<td>56.25</td>
<td>51.28</td>
<td>45.54</td>
<td>41.70</td>
</tr>
</tbody>
</table>

Table 1: Quantitative precision results at different recall levels for the nine BBC news broadcasts.

In order to reduce the computational complexity, in the approach solely relying on the face tracks and their similarity we obtain up to more than 86% precision at 70% recall. When combining the face track and the sparse matrix approaches, the results hardly decline. In this case, we set for each face the number of nearest neighbors $k_{f} = 90$ and for each face track, the number of nearest neighbors $k_{t} = 60$. In the latter setting memory usage and processing time are almost 15 times reduced.

In a final experiment, we test the portability of labeled faces across broadcasts. The label propagation framework perfectly allows the integration of labeled faces in the transition probability matrix $T^{*}$, which were annotated for another broadcast. In this way labeled faces can be reused across broadcasts. We have shown that in a particular setting, e.g. when using anchor detection and training data imported from a different broadcast (a very realistic setting), we manage to name 70% of the faces in the news videos with a precision of 85%.

4. REFERENCES
