Semi-supervised Identification of Rarely Appearing Persons in Video by Correcting Weak Labels

Eric Müller[1,2]
eric.mueller@tib.eu
[1]German National Library of Science and Technology (TIB)
Hannover, Germany

Christian Otto[2]
christian.otto@eah-jena.de
[2]University of Applied Sciences Jena
Jena, Germany

Ralph Ewerth[1,3]
ralph.ewerth@tib.eu
[3]Leibniz Universität Hannover
Hannover, Germany

ABSTRACT

Some recent approaches for character identification in movies and TV broadcasts are realized in a semi-supervised manner by assigning transcripts and/or subtitles to the speakers. However, the labels obtained in this way achieve only an accuracy of $80\% - 90\%$ and the number of training examples for the different actors is unevenly distributed. In this paper, we propose a novel approach for person identification in video by correcting and extending the training data with reliable predictions to reduce the number of annotation errors. Furthermore, the intra-class diversity of rarely speaking characters is enhanced. To address the imbalance of training data per person, we suggest two complementary prediction scores. These scores are also used to recognize whether or not a face track belongs to a (supporting) character whose identity does not appear in the transcript etc. Experimental results demonstrate the feasibility of the proposed approach, outperforming the current state of the art.

Keywords

Face identification in video, semi-supervised learning.

1. INTRODUCTION

The challenging task of person identification in movies and TV broadcast is fundamental for a broad variety of multimedia applications, such as face retrieval, video recommendation, and automatic character labeling. While some approaches rely on manually annotated training data [7], many other proposals [4, 9, 1, 6] employ semi-supervised learning where the labels for training are directly extracted from the test video. For that purpose, it is necessary to link a speaker from subtitles and/or transcripts to the visual content of the current face tracks. This scenario bears a couple of challenges since the required speaker detection is a difficult problem itself. As a result, the accuracy of the extracted *weak labels* is typically between 80% and 90%. Besides, the gallery of persons obtained this way is often highly imbalanced with respect to the number of examples per individual. This is caused by the uneven speaker distribution in a video, where some face tracks even remain unlabeled. Hence, a special treatment of these *unknown* persons is required.

In this paper, we present a semi-supervised system for person identification in video that addresses these issues. An initial prediction step is executed to find highly reliable results by combining a k-nearest neighbor approach with a complementary score that considers the distribution of similarities. The highly reliable predicted tracks are added to the training dataset, which increases the amount of training data of rarely appearing persons. In a next step, this additional knowledge is used to correct the original set of weak labels in order to enhance their accuracy. Furthermore, a combined criterion is utilized to identify face tracks of (unknown) persons that do not belong to an individual of the gallery set, e.g., supporting actors. Experimental results show that the proposed approach noticeably reduces the error rate of character identification in video.

The remainder of the paper is organized as follows. Section 2 discusses recent advancements in the field of person recognition in video. The proposed approach is described in detail in section 3. Section 4 reports experimental results, while section 5 concludes the paper and outlines some areas for future work.

2. RELATED WORK

Person identification in video can be separated into two categories depending on the source of knowledge concerning the target characters. Supervised approaches, such as Ortiz et al. [7], rely on manually annotated training data for each identity. Even though they yield promising results, fully-automatic, semi-supervised approaches as introduced by Everingham et al. [4] are often more desirable. The authors suggest a method that combines textual and visual cues to generate so called weak labels for frontal face images, which has been extended by Sivic et al. [8] to half-profile and profile views. Cinbis et al. [3] exploit frames of the same face track to model intra-person variations and sequences of different persons acting together in a shot for extra-class relations. Tapaswi et al. [9] argue that constraints such as "the same person cannot appear twice in one frame" are very helpful for person identification. To make these rules more viable, they add clothing features in case that the face is not visible in the current frame. In a similar way, Bäuml et al. [1] model the appearances of characters in a multinomial logistic regression (MLR).

ICMR'16 June 06-09, 2016, New York, NY, USA

ACM ISBN 978-1-4503-4359-6/16/06.
DOI: http://dx.doi.org/10.1145/2911996.2912073


One main drawback of all aforementioned approaches is the usage of partially incorrect weak labels for classification. A detailed discussion of this problem is presented by Tapaswi et al. [10], incorporating both positive and negative constraints between and within tracks to enhance speaker assignment. In addition, the weakly labeled data is unevenly distributed for the different characters. To conquer this imbalance, Cherniavsky et al. [2] iteratively add the top 10% of their recognition results to the training data in order to learn facial attributes. The issue of identifying face tracks from *unknown* characters is addressed by Ortiz et al. [7] using classification based on sparse representation.

3. PERSON IDENTIFICATION IN VIDEO

The algorithm proposed in this paper follows a similar strategy as the approaches discussed above, but without adding any external data, constraints, or clothing features.

The input of our person recognition system are face tracks extracted from the given video and labels $\mathcal{Y} = \{1, \ldots, K\}$ generated by assigning one of $|\mathcal{Y}| = K$ different possible speakers based on lip movement detection. Face sequences labeled with an assigned speaker are called *supervised tracks* and are used for training, the remainders are called *unsupervised tracks*. All face tracks are represented by the features described in section 3.1. Based on the prediction step introduced in section 3.2, the original set of supervised tracks is extended (section 3.3) and corrected (section 3.4). Then, face tracks of unknown persons are recognized based on a criterion introduced in section 3.5. Finally, an identity is predicted for each given face track (supervised as well as unsupervised) utilizing the improved input data.

3.1 Feature Extraction

In a first step, face images are aligned according to [5] and cropped to a size of 100×100 pixels. Subsequently, they are represented by the descriptor of Vu [11] combining patterns of oriented edge magnitudes (POEM) with patterns of orientation difference (POD). The representations are further divided into 8×8 blocks for a histogram representation using local binary patterns (LBP). A maximum of $L = 30$ time-sampled faces per track are added to the dataset. In order to reduce dimensionality and to select the most prominent features, principal component analysis (PCA) is applied followed by a whitening process to improve the robustness and discriminative power as proved by [11]. The latter one is learned a priori on the aligned *Labeled Faces in the Wild* dataset provided by [12]. Each face track t consists of a feature vector set $\mathcal{F}_t = \{f_i\}_{i=1}^{L}$. They are used to generate the supervised set $\mathcal{X}_s = \{(\mathcal{F}, y) \,|\, y \; corresponds \; to \; \mathcal{F}\}$ of N speaking tracks with the corresponding labels y and an unsupervised set $\mathcal{X}_u = \{(\mathcal{F}, \emptyset)\}$ of M non-speaker tracks.

3.2 Prediction with Complementary Scores

Two feature vectors f_1 and f_2 are compared using the cosine similarity, which is measured by:

$$s_{cos}\left(f_1, f_2\right) = \frac{f_1 f_2}{\|f_1\| \, \|f_2\|}. \tag{1}$$

The similarity between two track feature sets $s\left(\mathcal{F}_1, \mathcal{F}_2\right)$ is defined as the maximum of all (30×30 at most) scores of feature vector comparisons:

$$s\left(\mathcal{F}_1, \mathcal{F}_2\right) = \max\left\{s_{cos}\left(f_1, f_2\right) \mid f_1 \in \mathcal{F}_1 \wedge f_2 \in \mathcal{F}_2\right\}. \tag{2}$$

In the next step, the similarities of a track t to all N supervised tracks are calculated, sorted in descending order, and then stored in $\mathcal{S}_t$ along with their associated labels:

$$\mathcal{S}_t = \{(s(\mathcal{F}_t, \mathcal{F}_x)\,,\, y) \,|\, x \in \mathcal{X}_s \;\wedge\; y \; corresponds \; to \; \mathcal{F}_x\}. \tag{3}$$

This set is used as the input for the prediction step. Two different scores are suggested to recognize the identity $y_t \in \mathcal{Y}$ of track t. The first one determines the mean of the k highest similarity values regarding a certain character c:

$$p_t^{knn}\left(c\right) = \frac{1}{k} \sum_{i=1}^{k} s_i \;, s_i \in \mathcal{S}_t, y_{s_i} = c, k \in \mathbb{N}. \tag{4}$$

This is problematic if the currently compared character c has a low number of training samples, due to the higher variance of the k similarity scores. The lack of diversity, regarding the intra-class variations (e.g., pose, illumination, expression), will most likely lead to a low similarity score. This effect can be reduced by choosing a low value for k. There are two drawbacks of this approach: 1.) Outliers (in particular, due to incorrect track labels) skew the result heavily, and 2.) overfitting, since actors with a large number of training examples have a higher chance to match the pose and expression conditions of the current face track.

We address this problem by introducing a second score which considers all similarities. Therefore, the normalized similarities $\overline{s}$ are quantized into Q segments. To build a score, a weighting system is added, which rates each segment $w \in \mathbb{N}_{>1}$ times higher than its predecessor. The final cumulative score is calculated via the following equation:

$$p_t^{dis}(c) = \frac{1}{N_c} \sum_{i=0}^{Q-1} w^i \left| \left\{ \overline{s} \mid \overline{s} > \frac{i}{Q} \wedge y_{\overline{s}} = c, \overline{s} \in \mathcal{S}_t \right\} \right|, \tag{5}$$

where N_c denotes the number of training images of character c and $y_{\overline{s}}$ is the corresponding label for $\overline{s}$ according to equation 3.

The final prediction score of character c is calculated using the normalized result of equation (4) and (5). The $|\mathcal{Y}|$ final scores can be also exploited in order to decide whether track t shows a known or an unknown person, respectively. The decision is made using the criteria introduced in section 3.5. If it is a known person, the final prediction label y_t is determined by equation 6:

$$y_t = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \{\alpha \times \overline{p_t^{knn}(c)} + (1 - \alpha) \times \overline{p_t^{dis}(c)}\}, \tag{6}$$

with $\alpha \in [0, 1]$ weighting the two scores. Overfitting of persons with a large number of training examples is avoided by evaluating the distribution of similarities and taking all possible intra-class variations into account. We argue that these two scores complement each other well, which leads to a better identification performance.

Regardless, if the ratio between the training data of different characters is extremely large, a correct prediction is very difficult. Hence, we introduce a procedure that allows us to extend the number of supervised tracks for all persons in a reliable manner.

3.3 Adding Training Data

Poor recognition performance can be observed for persons with fewer appearances or speaking roles due to the resulting uneven distribution of training data. For this reason, the weakly labeled data $\mathcal{X}_s$ and an initial prediction process (as

described in 3.2) are employed to predict the unsupervised data $\mathcal{X}_u$. In order to only add reliable results, an additional ratio R is calculated for each track t, defined by dividing the second highest by the highest score according to equation (6). If this ratio fulfills the following confidence criterion;

$$1 - R \geq \tau_c, \tag{7}$$

with τ_c as a confidence threshold, then the face track and the corresponding predicted label is added to the (supervised) training data. These *extended supervised tracks* $\mathcal{X}_e$ not only enhance the prediction performance, but also open up the possibility for a robust *correction* of weak labels. This is essential to reduce the influence of outliers to the k-nearest neighbor classifier described in section 3.2.

3.4 Correcting Speaker Labels

As stated before (see also Bäuml et al. [1]), the assigned labels relying on speaker detection are typically noisy and therefore impair the recognition results. To tackle this problem, the extracted extended data $\mathcal{X}_e$ are used as a training set to correct them. For this purpose, another prediction process is conducted for each track in the original data $\mathcal{X}_s$ to predict a refined label y_t, replacing the original one. This step yields the *final supervised training dataset* $\mathcal{X}_f$.

3.5 Recognition of Unknown Persons

Bäuml et al. [1] annotate all background and supporting actors, detected during the speaker assignment process without an unique label in the subtitles or transcripts, as *unknown*. As a result, training data for an unknown person (and its related label) represents a variety of different persons, whereas the other labels always only denote exactly one individual. In contrast to Bäuml et al. [1], we claim that this variety makes it very challenging to assign the label *unknown* to the face track under consideration.

Based on this observation, we combine two conditions to detect if a track does not belong to a known identity ($y_t \notin \mathcal{Y}$). The first condition checks if the averaged k-nearest neighbor score from equation (4) is below a threshold τ_{knn}, while the second one checks if equation (7) is false. If both conditions are fulfilled, then the current face track t is labeled as $y_t = unknown$.

4. EXPERIMENTS

4.1 Experimental Setup

The dataset[1] of Bäuml et al. [1] is used in our experiments. It consists of the first six full episodes from season one of the TV sitcom *The Big Bang Theory* (*BBT-1* to *BBT-6*). The dataset contains 3,759 face tracks with a total of 11 identities and an additional label for unknown characters. In addition, 986 speaker labels are provided with an accuracy of 88.03%. The number of speaking tracks per character is shown in Table 1.

Our system is evaluated using the accuracy of identification according to [1], i.e., all face tracks are required to be labeled correctly, even if there are no training data available for an individual. The proposed approach is compared with Multinomial Logistic Regression + Markov Random Fields (MLR+MRF) of [1] and Mean Sequence Sparse Representation-based Classification (MSSRC) + Affinity of [6]. To the best of our knowledge, the latter one is the best result on BBT data. It should be noted that Bäuml et al. [1] exploit additional clothing features and Ortiz [6] uses only the supervised images of the current investigated episode, except for *Raj* and *Howard*, where training data from all episodes are taken into account to handle the character imbalance.

[1]From: http://cvhci.anthropomatik.kit.edu/projects/mma

Table 1: Number of training tracks N_c and the corresponding identification accuracy *with respect to all 3,759 tracks*, per character c for the BBT dataset.

	#training tracks		accuracy [%]	
	original	final	original	final
Leonard	300	966	95.23	97.29
Sheldon	320	791	94.81	97.78
Penny	170	424	88.09	92.38
Howard	73	225	84.28	86.29
Raj	32	109	50.54	67.74
Mary	49	85	85.26	80.00
Leslie	11	47	57.14	80.95
Kurt	14	33	81.25	87.50
Gabelhauser	4	4	0.00	6.25
Doug	0	0	0.00	0.00
Summer	0	0	0.00	0.00
Unknown	13	13	71.33	75.90
All	986	2,697	—	—

Table 2: Number of training tracks and their corresponding label accuracy for the BBT dataset.

	#training tracks	accuracy [%]
original[1] $\mathcal{X}_s$	986	88.03
extended $\mathcal{X}_e$	2,697 (986 + 1,711)	93.59 (88.03 + 96.79)
final $\mathcal{X}_f$	2,697 (986 + 1,711)	96.52 (96.04 + 96.79)

In the experiments, the prediction step described in 3.2 is applied to all 3,759 face tracks (supervised as well as unsupervised tracks) of the BBT dataset. We estimate the parameters for the experiments using $\mathcal{X}_s$ by setting all tracks of *howard* in $\mathcal{X}_s$ to *unknown*. Apart from requiring the parameters to yield the maximum accuracy for the remaining known characters, an auxiliary condition of a minimum prediction rate of 50% for *unknown* tracks has to be fulfilled. The estimated parameter setting is as follows: $k = 4$, $Q = 10$, $w = 2$, $\alpha = 0.8$, $\tau_c = 0.1$ and $\tau_{knn} = 0.71$. Since solely reliable tracks should be added in the extension step, τ_c is doubled to 0.20. Consequently, the parameters are re-estimated as described before, but assigning *howard* in $\mathcal{X}_f$ to *unknown* and requiring a minimum prediction rate of 75% for *unknown*, because the final data set contains more information and less incorrect labels. In this way, the parameter setting is obtained for the experiments with $\mathcal{X}_e$ and $\mathcal{X}_f$: $k = 3$, $Q = 10$, $w = 2$, $\alpha = 0.8$, $\tau_c = 0.2$ and $\tau_{knn} = 0.72$.

4.2 Results

The number of supervised face tracks and their related accuracy are displayed in Table 2. The extension of the training dataset improves the accuracy by 8.0% with respect to the original weak labels. The overall accuracy of the final label set X_f exceeds the baseline's accuracy by 8.5%. Besides, a total of 1711 tracks are additionally considered as training examples, which is spread over the different actors as shown

Table 3: Identification accuracy [%] for the BBT dataset.

	BBT-1	*BBT-2*	*BBT-3*	*BBT-4*	*BBT-5*	*BBT-6*	**AVG**
MLR+MRF [1]	95.18	94.16	77.81	79.35	79.93	75.85	**83.71**
MSSRC + Affinity [6]	95.19	90.53	86.00	84.21	83.11	85.91	**87.49**
Ours (original labels)	92.44	91.15	85.81	83.13	85.84	77.07	**85.91**
Ours (extended labels)	94.37	93.45	87.77	85.89	91.04	81.22	**88.96**
Ours (final labels)	96.30	94.87	89.23	87.09	92.47	81.59	**90.26**

in Table 1. It is obvious that the number of training tracks increases for each person except for *unknown* (which is intended) and for *Gabelhauser*. This can be explained by the very low number of four supervised tracks for *Gabelhauser* in the original training dataset, where only two of them have sufficient quality making it difficult to find additional training data. As a consequence, a prediction rate of only 6.3% for *Gabelhauser* is achieved.

Due to the higher number of training tracks, a gain in the identification accuracy can be observed for each known person (except mary) in Table 1. In particular, significant improvements for the characters *Raj*, *Leslie* and *Kurt* with less than 50 training tracks are achieved. It should be noted that no training material is available in the dataset for *Doug* and *Summer*, which explains the accuracy of 0%. Compared to the result of 13% reported by [1], the accuracy of predicting unknown persons is significantly increased to 75.9%, which is similar to the required value in the validation step.

The proposed system outperforms the results reported by Bäuml et al. [1], as displayed in Table 3, in each episode resulting in a higher prediction accuracy of 6.6% on the average. While only a slight improvement is achieved on *BBT-1* and *BBT-2*, more significant improvements can be observed on the remaining episodes. The reason is the higher number of *unknown* character occurrences in *BBT-3* to *BBT-6*, which proves the robustness of our treatment of *unknown* persons. Compared to Ortiz' system [6], the approach presented in this paper noticeably reduces the error rate from 12.5% to 9.7%, which is a relative improvement of 22.2%.

5. CONCLUSIONS

In this paper, we have presented a novel semi-supervised approach for automatically naming characters in TV broadcasts by extending and correcting weakly labeled training data. For this purpose, the combination of two complementary scores has been exploited. The common problem of an imbalanced number of training data and the handling of *unknown* persons have been addressed by the proposed approach as well. Experimental results have been reported on a benchmark test set of six episodes of the TV sitcom *The Big Bang Theory*. The system has achieved superior accuracy compared with two state of the art systems.

In future work, we plan to further enhance the identification process by weighting the scores depending on the number of training images per person. Another possibility for future research is to extract features using deep convolutional neural networks and to model them in a Joint Bayesian approach.

6. ACKNOWLEDGMENT

Part of this work is financially supported by the German Federal Ministry of Economic Affairs and Energy (BMWi, ZIM-KOOP, grant no. KF2135608KM3).

7. REFERENCES

[1] M. Bäuml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3602–3609. IEEE, 2013.

[2] N. Cherniavsky, I. Laptev, J. Sivic, and A. Zisserman. Semi-supervised learning of facial attributes in video. In *Trends and Topics in Computer Vision*, pages 43–56. Springer, 2012.

[3] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *IEEE International Conference on Computer Vision*, pages 1559–1566. IEEE, 2011.

[4] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy"–Automatic Naming of Characters in TV Video. In *British Machine Vision Conference*, pages 92.1–92.10, 2006.

[5] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874. IEEE, 2014.

[6] E. G. Ortiz. *Taming Wild Faces: Web-Scale, Open-Universe Face Identification in Still and Video Imagery*. PhD thesis, University of Central Florida Orlando, Florida, 2014.

[7] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3531–3538. IEEE, 2013.

[8] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?"-Learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1152. IEEE, 2009.

[9] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" probabilistic person identification in TV-series. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2658–2665. IEEE, 2012.

[10] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Improved weak labels using contextual cues for person identification in videos. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, volume 1, pages 1–8. IEEE, 2015.

[11] N.-S. Vu. Exploring patterns of gradient orientations and magnitudes for face recognition. *IEEE Trans. on Inform. Forensics and Security*, 8(2):295–304, 2013.

[12] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, 2011.