AUDIOVISUAL CELEBRITY RECOGNITION IN UNCONSTRAINED WEB VIDEOS

Mehmet Emre Sargin[*]*, Hrishikesh Aradhye, Pedro J. Moreno and Ming Zhao*

Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043

ABSTRACT

The number of video clips available online is growing at a tremendous pace. Conventionally, user-supplied metadata text, such as the title of the video and a set of keywords, has been the only source of indexing information for user-uploaded videos. Automated extraction of video content for unconstrained and large scale video databases is a challenging and yet unsolved problem. In this paper, we present an audiovisual celebrity recognition system towards automatic tagging of unconstrained web videos. Prior work on audiovisual person recognition relied on the fact that the person in the video is speaking and the features extracted from audio and visual domain are associated with each other throughout the video. However, this assumption is not valid on unconstrained web videos. Proposed method finds the audiovisual mapping and hence improve upon the association assumption. Considering the scale of the application, all pieces of the system are trained automatically without any human supervision. We present the results on 26,000 videos and show the effectiveness of the method per-celebrity basis.

Index Terms— Speaker recognition, Face recognition.

1. INTRODUCTION AND PRIOR WORK

Unobtrusive person recognition has been studied extensively using both face [1] and voice-based [2] biometrics. Existing approaches for conversational speaker recognition have mostly focused on telephonic domain, where only the audio modality is available. In the video domain, approaches that use both voice and face modalities have been shown to outperform unimodal methods in noisy environments [3], [4]. The principal concept in these approaches is to make use of the modality that is less effected by noise, thereby improving system performance. Many of the existing audiovisual person recognition systems, unfortunately, assume a tightly controlled data-capture environment and a cooperative subject. In the training phase, each subject is requested to record a known pass phrase, which is then matched with a probe phrase for authentication purposes. This scenario is referred to as *text-dependent* person recognition. The reader is referred to [5, 6] and some of our prior work in this area [7].

The scenario of interest to this work requires text-independent recognition since the content, quality, and capture environment of web videos is completely unconstrained, even for the videos involving celebrities. To the best of our knowledge, this problem has been addressed in the published literature only within the constrained subset of anchorperson recognition in broadcast news [8]. News anchors appear in controlled illumination often in a *talking head* view with a stationary camera, and often read scripted monologue off the teleprompter in (relatively) long, uninterrupted segments. Furthermore, the authors considered only those clips with frontal facial views. The problem at hand allows for no such assumptions.

In our previously published work [9], we presented a method for recognizing celebrity faces in unconstrained web videos. Our method differed from the rest of the face recognition literature primarily in its ability to train autonomously by learning a *consistent* association of faces detected in an image on a webpage with person names found in the text of the webpage. The internet is in a constant state of flux, and new "celebrities" are constantly added to the popular culture even as the celebrities of the past fade. This ability to learn autonomously to constantly add to the existing gallery of celebrities is therefore a major design principle of our work. Our face-based celebrity recognition system can recognize hundreds of thousands of celebrity faces at this point by exploiting the tremendous depth of the internet with the consistency learning framework. However, unimodal recognition based solely on faces is hampered when the image quality is poor and/or when facial details are blurred due to motion or occlusion. The proposed method is a logical extension of our existing face recognition system to exploit the biometric characteristics of the voice modality.

Continuing in the spirit of autonomous training, our method does not need any explicit manual labeling. It learns by automatically learning a consistent association of voice signatures with recognized faces and text from video title and tags. In doing so, it discards most cases where the face-voice association is inconsistent, most commonly with slideshow

[*]M.E. Sargin (msargin@ece.ucsb.edu) is now with Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 93106.

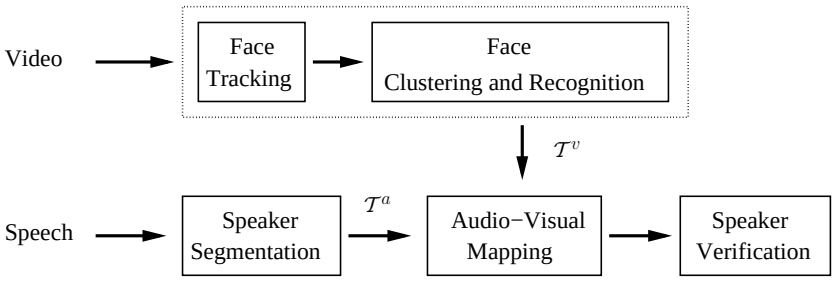

Fig. 1. System Overview. The experimental results are obtained through the speaker verification with and without using the audio-visual mapping.

videos (photos of celebrities combined with music), broadcast news (anchorperson talking about the celebrity with his/her photograph on display) and talk shows (camera focusing on one of the celebrities while the other person is talking).

2. PROPOSED APPROACH

Proposed system consists of four main blocks as illustrated in Figure 1. Face tracking followed by face recognition yields cohesive segments in time and space where the same (or similar) face is seen, assigning the same label even when the face is seen in temporally disjoint segments in the video. Similarly, speaker segmentation results in segments of the audio track where voice signatures are similar, assigning the same label even when the voice is heard in temporally disjoint segments. The set of time indices from both audio and visual tracks ($\mathcal{T}^a$ and $\mathcal{T}^v$, respectively) are then combined to find consistent one-to-one associations between the face tracks and the speaker segments. Consequently, for each recognized face, the biometric characteristics of voice from the corresponding speaker segments are used for speaker verification.

We now provide a detailed description about the components of the proposed system. Person segmentation in speech and visual domain are explained in Section 2.1 and Section 2.2 respectively. The fusion of audiovisual label sequences is described in Section 2.3. The learning and verification subtasks for speaker verification are explained in Section 2.4 and Section 2.5 respectively.

2.1. Audio Representation

We characterize the audio signal by 13 Mel-Frequency Cepstral Coefficients (MFCC) together with first and second order derivatives. These 39 dimensional feature vectors have a frame rate of 100 frames per second. The audio signal is first segmented into speech vs. non-speech using a Finite State Transducer (FST). Each state (speech, music, silence and noise) emits observations based on a GMM. Although the FST was trained using telephonic data, its segmentation performance was observed to be robust to channel variations when applied to unconstrained videos.

Agglomerative clustering is then applied to the speech segments. Distances between each cluster are calculated by a modified version of Bayesian Information Criterion (BIC) [10]. Densities within each cluster are approximated using a single Gaussian with a full covariance matrix. We assume that only one speaker is speaking at any given time.

Let us assume that the number of unique speakers in the entire audio track is represented by S. Here, S is upper bounded by the number of speech segments in the audio. For each speaker s, $s \in \{1, 2, \ldots, S\}$, we construct a set of time indices, $\mathcal{T}_s^a$, corresponding to the speech segments of s.

2.2. Visual Representation

Face detection and tracking are applied on the video as described in [9]. Face tracks are clustered using a methodology similar to the one described in Section 2.1 for speaker segmentation. A set of key faces are extracted from each cluster of face tracks which is subjected to face-based celebrity recognition. We refer the reader to our previous work [9] for details of the face recognition system.

Let us assume that the number of unique faces in the entire video is represented by F. Here, F is upper bounded by the number of face tracks in the video. For each face f, $f \in \{1, 2, \ldots, F\}$, we construct a set of time indices, $\mathcal{T}_f^v$, corresponding to the face tracks of f. Multiple faces may be present on the same image, hence some of the time indices may be included in multiple $\mathcal{T}^v$s.

2.3. Audiovisual Mapping

Ideally, S and F would be the same. In unconstrained web videos, even with perfect audio and video segmentations, it is still possible to have cases where S is not equal to F due to voiceovers, split-screens, and camera selection.

Joint probability of a pair of audio and visual labels belonging to the same person is estimated using the following

formula:

$$P(s,f) = \frac{1}{\alpha}\left|\mathcal{T}_s^a \cup \mathcal{T}_f^v\right|. \quad (1)$$

Here, α is the normalization parameter so that $\sum_{s,f} P(s,f) = 1$. Based on $P(s,f)$, we are interested in finding K ($K \leq \min(S,F)$) one-to-one association pairs in such a way that the joint probabilities of all pairs $\mathcal{M} = \{(s_k, f_k)\}$ are greater than an acceptable threshold $\theta_p > 0$. We used the following greedy algorithm to obtain one such mapping.

Algorithm 1 Audio-Visual Mapping

$\mathcal{M} \Leftarrow \{\}$
for $k = 1$ to $\min(S,F)$ **do**
  $(s^*, f^*) \Leftarrow \underset{(s,f)}{\operatorname{argmax}} P(s,f)$
  if $P(s^*, f^*) > \theta_p$ **then**
    add (s^*, f^*) to $\mathcal{M}$
    $P(s^*, :), P(:, f^*) \Leftarrow 0$
  else
    return $\mathcal{M}$
  end if
end for
return $\mathcal{M}$

2.4. Learning

We constructed a Universal Background Model (UBM) as a GMM using speech segments that are not associated with the celebrities of interest. The UBM is used as a starting point for celebrity model estimation as well as as a null hypothesis for speaker verification. We used 1024 mixtures of Gaussians with diagonal covariances with standard maximum likelihood GMM training. A GMM for each celebrity is obtained by MAP adaptation using the UBM as the prior. During the adaptation process we only updated the means [11].

2.5. Verification

Let $\mathbf{X} = \{x_t\}, t \in \mathcal{T}_c^a$ represent the MFCC feature vectors extracted from a video where $\mathcal{T}_c^a$ is the set of time indices corresponding to the speech segments of the hypothetical celebrity c extracted from audio-visual mapping. We either accept or reject the hypothesis of $\mathbf{X}$ being associated with celebrity c by the following formula:

$$\frac{1}{|\mathcal{T}_c^a|}\sum_{t\in\mathcal{T}_c^a}\{\log p(x_t|c) - \log p(x_t|UBM)\} \underset{reject}{\overset{accept}{\gtrless}} \theta. \quad (2)$$

3. EXPERIMENTAL RESULTS

3.1. Training and Testing Data

Our experimental dataset consists of 4M most popular YouTube videos. A total of 2600 celebrities were recognized by the face recognition system in 730K of these 4M videos. Consistent with our underlying design principle of scalability and automatic learning, we avoided any manual annotation of the dataset. User-supplied title and keywords represent one source of ground truth annotation. However, presence of celebrity names in user-suppled metadata does not necessarily imply the presence of those celebrities in the video footage, and conversely, the lack of such names in the metadata does not necessarily imply that the celebrities are not present in the video footage. A significant subset of the videos does not have any user-supplied keywords at all. Face tracking and recognition results provide another source of annotation, which has its own imperfections. To train and test voice models for celebrities of interest, we selected only those videos where face-based celebrity hypothesis agreed with user-supplied metadata – a set of 26K videos with 200 celebrities. Note that this "ground truth" data may still have incorrect labels, such as (1) the celebrity in question may not be speaking during part or whole of the face track, (2) errors in face track clustering may incorrectly group distinct individuals into one identity. Such imperfections in the ground truth, along with the sheer size and unconstrained nature make this a challenging dataset. However, this procedure can be carried out completely autonomously for any new celebrities rising in the popular culture as reflected on YouTube, especially given the fact that the face recognition system is also trained completely automatically.

3.2. Verification Performance

Randomly selected two thirds of the 26K videos are used as training. The rest are used for testing. Imposter videos are selected randomly from the test data (excluding the ones that have celebrity of interest) with the amount proportional to the actual videos of the celebrity of interest. We obtained false accept and false reject rates by changing θ in Equation 2. The Equal Error Rates (EER) for each celebrity are presented in Figure 2. Two different configurations of the proposed system are tested. Results for the first configuration (dashed line in Figure 2) are obtained without considering the audiovisual mapping block. In this case, all speech segments corresponding to the time spans for face tracks associated with hypothetical celebrity c, $\mathcal{T}_c^v$, are used for voice characterization. Alternatively, EER results with audiovisual mapping are shown as a solid line in Figure 2. Audiovisual mapping improved the median EER across celebrities by 5.5%.

An interesting result that can be inferred from Figure 2 is that EER for each celebrity is correlated with the "talkativeness" of that celebrity in web videos. Most of the celebrities that have EER $< 10\%$ (such as Michelle Malkin, Larry King and Bill O'Reilly) are popular talk show hosts. EERs for actors and actresses are seen to be lower, as they speak less often.

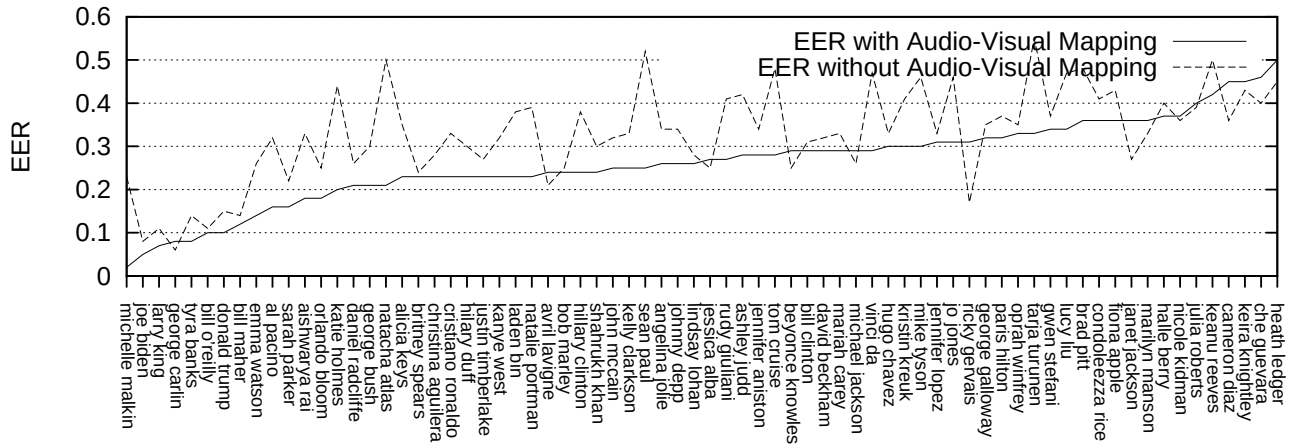

Fig. 2. EERs for each celebrity. Solid and dashed lines represent the EER with and without using the audiovisual mapping respectively. Celebrity names are sorted by their EERs using the audiovisual mapping for readability.

4. CONCLUDING REMARKS AND FUTURE WORK

In this paper, we present an audiovisual celebrity recognition system that integrates face-based recognition and voice-based verification modules at the decision level. The results presented in this paper, while not as good as the state of the art speaker verification systems, are very encouraging since the underlying domain (large scale videos on YouTube vs. telephonic speech) is far less constrained and the ground truth is imperfect. To the best of our knowledge, no such system exists in the published literature.

A large number of videos are uploaded on YouTube every day, and new celebrities constantly rise and fade in the popular culture. Conventional learning approaches that require manually annotated data will not scale well in the application scenario of interest to this work. Therefore, all recognition components of our system train autonomously without needing any manually labeled training data.

Unlike the controlled audio-visual data-sets, unconstrained manual editing often makes the audio and visual streams completely asynchronous, hence the need for investigation of audio-visual association on the web videos. To this end, we proposed a new algorithm for consistent association of face and voice segmentation subsystems. This audiovisual mapping was demonstrated to significantly improve the overall performance. Improvement in this mapping is being investigated as ongoing work via speaking face detection by lip motion estimation.

5. REFERENCES

[1] W. Zhao, R. Chellappa, PJ Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[2] A. Stolcke, S.S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on mllr adaptation transforms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 1987–1998, Sept. 2007.

[3] Tsuhan Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, Jan 2001.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept. 2003.

[5] Tieyan Fu, Xiao Xing Liu, Lu Hong Liang, Xiaobo Pi, and A.V. Nefian, "Audio-visual speaker identification using coupled hidden markov models," *IEEE ICIP*, vol. 3, pp. III–29–32 vol.2, Sept. 2003.

[6] A.V. Nefian and Lu Hong Liang, "Bayesian networks in multimodal speech recognition and speaker identification," *Signals, Systems and Computers, 2003. Asilomar Conference on*, vol. 2, pp. 2004–2008 Vol.2, Nov. 2003.

[7] M.E. Sargin, Y. Yemez, E. Erzin, and A.M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *Multimedia, IEEE Transactions on*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.

[8] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news: some fusion techniques," *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on*, pp. 161–167, 1999.

[9] Ming Zhao, Jay Yagnik, Hartwig Adam, and David Bau, "Large scale learning and recognition of faces in web videos," *Automatic Face and Gesture Recognition, 2008. FGR 2008. 8th Int. Conf. on*, September 2008.

[10] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, Aug. 2004.

[11] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.