Towards Person Google: Multimodal Person Search and Retrieval

Lutz Goldmann, Amjad Samour & Thomas Sikora*

Communication Systems Group
Technical University of Berlin
Berlin, Germany

Abstract. Content based multimedia retrieval systems have been proposed to allow for automatic and efficient indexing and retrieval of the increasing amount of audiovisual data (image, video and audio clips). The search for specific persons within this data is an important subtopic due to its large range of applications. This article describes an original system for multimodal person search and provides some initial performance results that demonstrate the efficiency of the system.

1 Introduction

With the increasing amount of available multimedia data, efficient systems for searching and retrieving relevant AV documents are needed. Since keyword based indexing is very time consuming and inefficient due to linguistic and semantic ambiguities, content based multimedia retrieval systems have been proposed, that search and retrieve AV documents based on audio and visual features. While content based image retrieval has been a very active research field, only some work has been done in the field of person search and retrieval, where the goal is to find a AV document with a specific person present within the audio and the visual stream. An original system for multimodal person search and retrieval is proposed in this article which is based on audio and video analysis techniques combined by a multimodal fusion approach.

2 System overview

Figure 1 gives an overview of the proposed system. The initial system is based on the query by example paradigm, where the user selects an AV document, the system compares it to the AV documents in the database and retrieves them ranked according to their similarity.

2.1 Audio analysis

The goal of the audio analyis part is to retrieve audio segments based on the voice characteristics of a person without considering the spoken content.

* The work presented in this paper was supported by the European Commission under contract FP6-027026 K-Space.

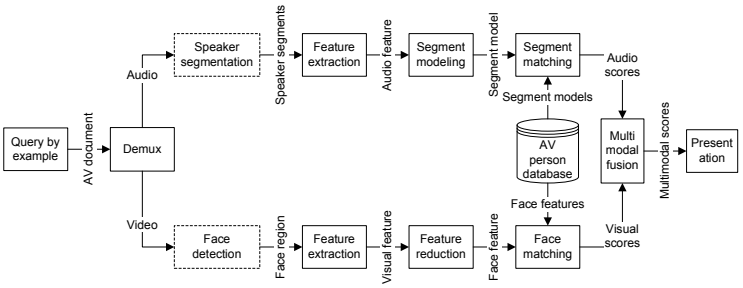


Fig. 1. Overview of the person search and retrieval system.

Feature extraction For describing the audio segments mel frequency cepstral coefficients (MFCC), introduced by Davis et al. [1], are applied. They are widely used in the area of speech recognition and audio classification since they provide a good description of audio characteristics under a wide range of conditions and with reasonable computational costs. A sliding window divides the audio segment into multiple overlapping frames with a length of 20 ms and an overlap of 10 ms. For each frame a feature vector consisting of 13 MFCC's and the log energy of the frame is created. Depending on the segment length, different number of feature vectors are extracted.

Segment modeling In order to reduce the temporal characteristics of the audio data within a segment and to create a robust model of the spectral characteristics of the speaker's voice, each speaker segment is modeled as a multivariate Gaussian distribution.

Segment matching The goal of the matching stage is to compare an audio segment with all audio segments within the database. Since each audio segment is described using a statistical model, model selection techniques are suitable for the comparison. The Bayesian information criterion (BIC), proposed by Schwarz et al. [2] is applied to compute the distance between two segments.

2.2 Visual analysis

The goal of the visual analysis part is to retrieve persons based on their facial appearance.

Feature extraction The face region is determined based on the pupil positions, obtained either manually or automatically, and an anthropometric face model [3]. In order to handle different face sizes, each face region is scaled to a common size. Since uneven illumination tends to change the appearance of faces, statistical normalization methods are applied globally and locally.

Feature reduction In order to reduce the feature dimensionality and maintain the most relevant information, the principal component analysis (PCA) [4] is applied. This leads to a set of eigenfaces that form a reduced basis onto which the original feature vectors are projected.

Feature matching The goal of the matching stage is to compare a query face with all faces in the database. Since each face is described by a feature vector, vector distances can be used for the comparison. For the initial experiments the Euclidean distance was chosen.

2.3 Multimodal fusion

The general goal of multimodal fusion is to exploit the complementary character of multimodal sources to increase the robustness of the system with regard to the single modalities. More specifically, the idea here is to combine the voice and face characteristics to resolve ambiguities within an individual modality. For the proposed system score level fusion was chosen since it provides the best tradeoff in terms of information content and ease of fusion.

Score normalization Scores of different modalities usually exhibit quite different characteristics (type, distribution, range) which make it very difficult to combine them in a suitable way. The goal of the score normalization step is to modify the location and variation of their distributions to transform them into a common domain. Out of the large number of possible normalization techniques the z-score normalization has been chosen for the initial system.

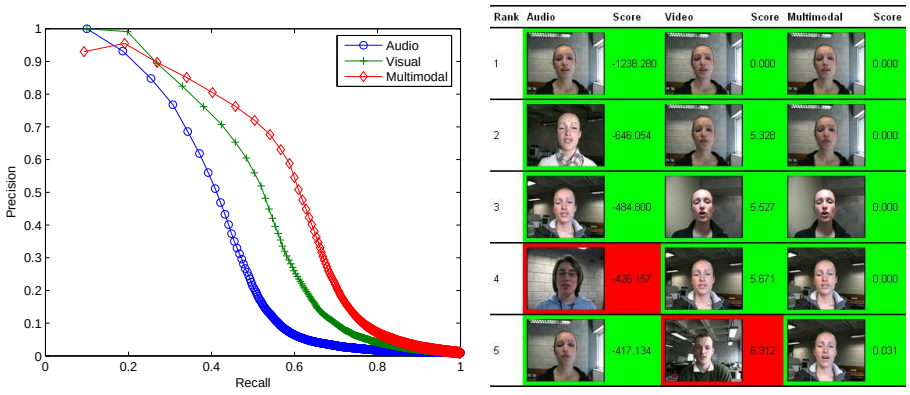
Score fusion Different fusion rules (product, sum, min, max) are considered within the system. For each of the AV documents, these fusion rules combine the corresponding audio and visual scores into a multimodal score.

3 Experiments

The initial experiments are based on the VALID database that contains multimodal data (audio, video) of 106 persons (27 female, 79 male).

Several evaluation measures have been proposed for evaluating search and retrieval systems [5]. They can be divided into precision/recall and rank measures. Both classes of measures have been considered.

Figure 2 shows a comparison of the audio only, visual only, and the multimodal system based on precision vs. recall (PR) curves and a sample query. In the current system, the visual modality outperforms the audio modality. Furthermore, it can be seen that the multimodal system shows a considerable performance improvement over the single modality systems. For 20 retrieved documents the mean recall values of the audio, visual, and multimodal system are $R = \{49; 58; 66\}\%$ while the corresponding precision values are $P = \{24; 29; 33\}\%$ respectively.



(a) Overall PR curve (b) Sample query

Fig. 2. Retrieval performance of the different variants (audio, visual, multimodal).

4 Conclusions

A system for multimodal person search and retrieval using voice and face characteristics has been developed. Initial experiments provide encouraging results and justify the proposed solution. While both modalities perform well individually, a performance improvement can be achieved with the multimodal system.

Future work will explore the different parts of the system in more detail and evaluate variants of the system. Furthermore, different query paradigms and relevance feedback techniques will be incorporated into the system. Another aspect is to analyze the influence of the detection steps (face detection, speaker segmentation) onto the retrieval performance.

References

1. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4) (1980) 357–366
2. Schwarz, G.: Estimation the dimension of a model. In: Ann. Stat. Volume 6. (1978) 461–464
3. Farkas, L.G.: Anthropometry of the Head and Face. Raven Press (1994)
4. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 4–37
5. Mueller, H., Mueller, W., et al.: Performance evaluation in content based image retrieval: Overview and proposals. Technical report, University of Geneva (1999)