Multimodal Person Search Combining Information Fusion and Relevance Feedback

Lutz Goldmann [*1], Amjad Samour [#2], Touradj Ebrahimi [*3], Thomas Sikora [#4]

[*] *Ecole Polytechnique Federale de Lausanne, Multimedia Signal Processing Group*
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
[1] lutz.goldmann@epfl.ch
[3] touradj.ebrahimi@epfl.ch

[#] *Technical University of Berlin, Communication Systems Group*
Einsteinufer 17, 10587 Berlin, Germany
[2] samour@nue.tu-berlin.de
[4] sikora@nue.tu-berlin.de

***Abstract*—With the increasing amount of multimedia data, efficient tools for search and retrieval are needed. Since people are naturally one of the most interesting objects within these documents, a system for multimodal person search and retrieval has been developed. It combines the audiovisual analysis of persons with the query by example paradigm and relevance feedback to provide an efficient tool for searching multimedia data. For the relevance feedback, one and two class approaches are considered and compared to each other. Multimodal fusion techniques are used to exploit the complementary character of the audio and video information. The experimental results prove that multimodal person search and retrieval is feasible and more efficient than manual exploration.**

I. Introduction

With the increasing amount of available multimedia data, efficient systems for searching and retrieving relevant audio-visual (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 multimedia 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 AV documents 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 the combination of the audiovisual analysis of persons with content based multimedia retrieval techniques such as query by example and relevance feedback.

The general idea is the following: given a large set of AV clips, containing individuals giving talks or delivering monologues, the goal is to find and retrieve all the clips of a specific person by providing a sample to the search engine. Typical application scenarios of such a system are illustrated in figure 1, including official video podcasts, personal video blogs and broadcast news. For most of these scenarios it can be assumed, that the voice present in the audio stream belongs to the person visible in the video stream.

Fig. 1. Application scenarios for multimodal person search and retrieval. From left to right: official video podcast, personal video blog, broadcast news.

As already mentioned very little work has been done in the field of multimodal person search and retrieval. Nevertheless, related work can be found in two major areas: *content based multimedia retrieval* and *multimodal biometrics*. The former deals with the search of multimedia documents usually without any emphasis on a certain object class [1]. The latter focuses on the identification of persons based on different biometric traits such as face, gait, voice, fingerprint and iris [2]. The most closely related work tries to search specific persons within images by combining keyword based search with face detection. This approach is used within Google's Image Search[1] and IDIAP's Google Portrait[2]. Another approach which is considered by Riya[3] is to combine user tagging and visual analysis of images to support search and retrieval of individuals.

In contrast to these approaches, the system described in this paper combines multimodal biometrics with content based retrieval techniques to retrieve humans within audiovisual sequences. It is an extension of the system described in [3] which apart from query by example considers also relevance feedback for improved retrieval performance. Both positive only and positive/negative relevance feedback are explored.

[1] http://images.google.com/
[2] http://www.idiap.ch/googleportrait/
[3] http://www.riya.yom/

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

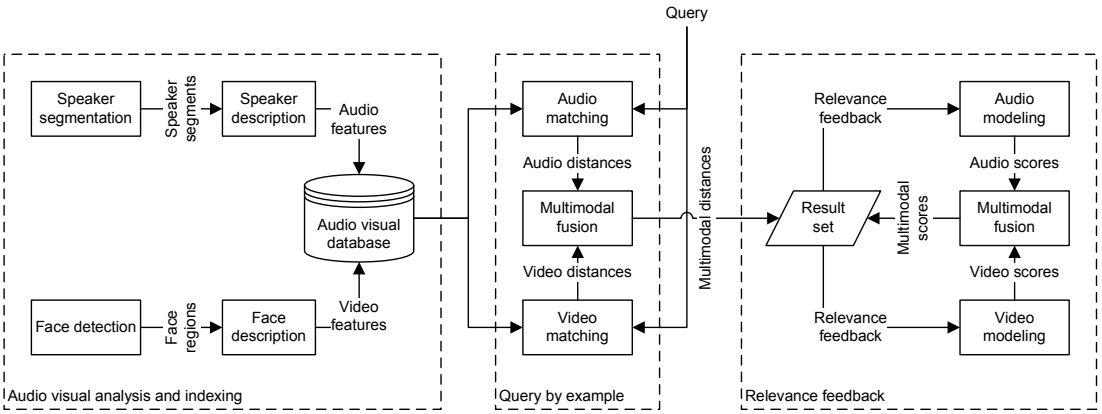

Fig. 2. Overview of the proposed system for multimodal person search and retrieval.

II. System overview

Figure 2 provides an overview of the proposed system for multimodal person search and retrieval. It consists of an offline and an online part. For a given database the *offline* part splits the multimodal data into an audio and a video stream, runs the audio and video analysis individually, and stores the corresponding information into a database. The *online* part itself consists of two parts, a *query by example* stage to start the search process and a *relevance feedback* loop, that is used to refine the search based on the feedback provided by the user. Again the audio and video information is treated individually during the *matching* and *modeling* steps, but finally combined within the multimodal fusion stage.

A. Audio analysis

The goal of the audio analysis part is to retrieve audio segments based on the voice characteristics of a person independent of the spoken content. It consists of an optional speaker segmentation step which segments the speech stream into individual speaker segments and a speaker description step that extract suitable audio features to describe each speaker's voice.

1) Speaker segmentation: The goal of the speaker segmentation step is to divide the audio stream into temporal segments corresponding to individual speakers. Therefore change points between different speakers are detected with a metric based segmentation approach similar to the one proposed by [4].

The audio stream is divided into frames of 40 ms duration for which well known Mel frequency cepstrum coefficients (MFCC) are computed. Given this sequence of audio features the Bayesian information criterion (BIC) is used to determine speaker changes. This is achieved by moving a sliding window over the stream and considering the corresponding features vectors as a single (whole window) or two individual Gaussian processes (two half windows). Then the decision if the window contains a change point or not can be interpreted as a model selection problem based on the BIC value. The temporal segment between two change points is then considered to belong to the same speaker.

2) Speaker description: The goal of the speaker description step is to extract a robust description of the speakers voice characteristics independent of the spoken content and environmental conditions. Again, well known MFCCs have been adopted since they provide a compact representation of the spectral characteristics of an audio signal that resembles the human auditory system.

Given a speaker segment MFCCs are extracted in the same way as in the speaker segmentation by dividing the audio stream into frames of 40 ms length, applying a Hamming window and computing the power spectrum. The power spectrum is transformed to the Mel scale by applying a set of triangular filters. Finally, the discrete cosine transform (DCT) is applied to compute the cepstral coefficients from the mel spectrum leading to a feature vector of size 13 for each window. In order to reduce the temporal characteristics of the spoken content within a segment and to create a robust model of the spectral characteristics of the speakers voice, each speaker segment is described by the arithmetic mean computed over all the windows.

B. Video analysis

The goal of the video analysis stage is to detect and describe the faces of present persons within the video.

1) Face detection: The first step within the visual analysis part is to detect the face of the present person within the individual frames of the video. In the current system the component based approach by Goldmann et al. [5] has been adopted. It has been shown that this approach can not only detect partially occluded faces, but also localizes these occlusions. This additional information can be used to select unoccluded samples of the persons face to increase the robustness of the retrieval process.

This face detection approach combines statistical and structural pattern recognition techniques. Facial components are detected using a combinaton of Haar features and an Adaboost trained classifier cascade. Possible combinations of the detected components are compared to a graph model of the face to accept or reject these facial candidates. Finally the

position of the detected face is estimated based on the detected components.

2) Face description: The face is described using the so called eigenface approach [6]. Therefore, detected faces are normalized and scaled to a common size (40x30 pixels) by applying geometrical transformations (scaling, translation, rotation) based on the pupil positions provided by the face detection step.

In order to handle uneven illumination a *local normalization* technique is applied to the extracted texture template. Therefore the overall region is split into a predefined number of subblocks (4x3 pixels) on which contrast stretching is applied individually. This approach considerably decreases the variation caused by uneven illuminations. Finally, a feature vector is extracted by row wise scanning the texture template.

Since the high dimensionality of the resulting feature vector may lead to the curse of dimensionality, feature reduction techniques are applied. Due to the unsupervised nature of the retrieval process, *principal component analysis (PCA)* instead of linear discriminant analysis (LDA) is applied. It finds a linear projection that maximizes the total scatter of the data. Finding the optimal projection basis is equivalent to computing the eigenvectors of the total scatter matrix. The corresponding eigenvalues provide a measure on how much variance each dimension contains. Thus the eigenvectors are sorted according to their eigenvalues and a subset of them is chosen as the reduced basis. The reduced feature vectors are obtained by projecting the original feature vectors onto the reduced basis.

C. Query by example (QBE)

The idea behind the query by example paradigm for retrieval is to ask the user for a sample that represents his search intention. This sample is analysed in the same way as all the samples within the database and compared to them based on some criteria. In the current system each sample is represented by two feature vectors, one for each modality. For each modality the distances between the corresponding sample and all the documents in the database are computed.

Several metrics have been proposed for the comparison of two feature vectors. The well known *Minkowski metric* which is basically a family of distance metrics has been chosen for this work. For two vectors x and y it is defined as

$$d_p(x,y) = \left(\sum_i (x_i - y_i)^p \right)^{1/p} \tag{1}$$

where p is the parameter that defines the different norms. After some initial experiments the *euclidean distance* ($p = 2$) has been chosen as the most appropriate metric.

D. Relevance feedback (RF)

Relevance feedback approaches can be categorized based on several criteria [7]. Out of the possible *time ratios* only current and previous rounds are considered. From the *different sources* only the information provided by the current user is considered. The relevance feedback process itself typically consists of a learning and a selection step. Within the *learning* step, a model for the user's search intention is built based on the provided feedback. From the proposed learning approaches (query point movement, reweighting, single Gaussians, support vector machines) the two latter have been selected, since they naturally support different types (one class and two class) of relevance feedback. After the learning step, a selection step is used to choose the items which are returned as the result set. Basically two approaches with very different goals can be used. Selecting the *most positive* items provides the user in each feedback round with a results set that shows him the best matches regarding his search intention. On the other hand selecting the *most informative* items tries to reduce the ambiguity and obtain feedback for critical decisions. Within the current system only the former selection criteria is used.

1) Single Gaussian (SG): The idea of the first relevance feedback strategy is to utilize only positive feedback to estimate the user's search intention [8]. It is assumed to follow a multivariate Gaussian distribution

$$p(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{2}$$

with the mean μ and the covariance matrix Σ. These parameters are estimated from the positive samples provided by the user. The matching is performed by computing the likelihood $p(x)$ for each sample in the database given the previously estimated model.

2) Support vector machine (SVM): The idea of the second relevance feedback strategy is to consider positive and negative feedback, by using a support vector machine (SVM) to discriminate between relevant and non relevant documents. The support vector machine (SVM) is a linear classifier that finds the optimal separating hyperplane

$$w^{\mathrm{T}}x + b = 0 \tag{3}$$

Out of the infinite number of possible hyperplanes the SVM looks for the one that maximizes the margin between the two different classes. In order to deal with non linear separation, the margin constraints can be relaxed by a penalty factor C. Furthermore, different kernel functions are used to map the original features into a higher dimensional space in which a linear separation is possible. From the set of commonly used kernels the radial basis function (RBF) kernel, defined as

$$k(x,z) = \exp(-\|x-z\|^2/2\sigma) \tag{4}$$

with the variance σ was chosen, since it has been shown that this kernel provides the best performance [9].

For matching the samples of the database to the trained SVM, distance from the decision boundary [9] is used. It is defined as

$$d(x) = \frac{\alpha k(x,z) + \beta}{\alpha z} \tag{5}$$

with the kernel function $k(x,z)$, the support vectors z, the scaling parameter α and the bias β.

E. Information fusion

The general goal of information fusion is to combine the information of different sources [10] and exploit this possibly uncorrelated information to improve the performance in comparison to the individual sources.

Approaches can be broadly categorized into two categories, depending on where the fusion is applied with respect to the mapping (e.g. classification) step [10]. While *premapping fusion* combines the original information before the mapping step, *postmapping fusion* integrates the information provided by the mapping step. In the current system postmapping fusion and more precisely score mapping fusion was considered.

1) Score normalization: Since the scores of the different modalities may have very different characteristics, a direct combination of them is not very reliable. Thus the location and scale of the different score distributions are modified to map them into a common domain. In the current system this is done independently for each RF iteration after the mapping.

The *z-score normalization* and its adaptation the *3-sigma normalization* have been proven to be quite reliable if the scores are following a Gaussian distribution. If the mean μ and the standard deviation σ of the scores is not known a priori, they are estimated from given samples. Using the 3-sigma normalization the scores are normalized according to the following equation

$$s' = \frac{s - \mu}{3\sigma} \tag{6}$$

which maps 99% of the scores into the range $[-1, 1]$. These scores are further transformed into the range $[0, 1]$ by shifting and clipping.

2) Score fusion: As already mentioned before, score level fusion has been chosen to combine the different modalities. Generally existing approaches are either classification or combination approaches. While the first group relies on post classifiers to reach a combined decision, the latter uses different combination rules to fuse the scores and a decision rule to reach the combined decision. The combination approach was chosen for this system. Several combination rules have been proposed [11].

Within this work four rules have been considered, that can be applied to various types of scores s (distances and probabilities). The *product rule* assumes statistical independence of the different modalities m. In general different biometric traits (face, voice) are mutually independent. The joint scores are given by $s = \prod_m s_m$. Apart from statistical independence the *sum rule* also assumes that the posterior probabilities do not deviate much from the prior probabilities. Thus it is applicable if a high level of noise leads to ambiguity in the classification problem. The joint scores are obtained by $s = \sum_m s_m$. The *min rule* is derived by bounding the product of posterior probabilities and computes the joint scores as the minimum $s = \min_m s_m$. The *max rule* approximates the mean of the posteriori probabilities and fuses the scores by taking the maximum: $s = \max_m s_m$.

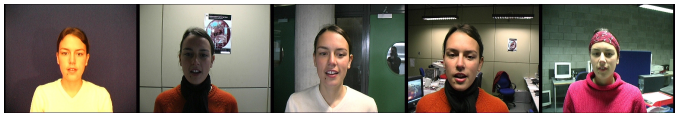

Fig. 3. Sample of the VALID database showing the different enviroments (studio, office).

III. Experiments

Several experiments have been conducted to assess the performance of the system and to identify the optimal combination of the different modalities. Furthermore, other aspects of such a system have been analysed, including

- Comparison of unimodal (audio, video) and multimodal approaches
- Comparison of different retrieval strategies including query by example (QBE) and relevance feedback (RF)
- Influence of different result set sizes into the retrieval performance and speed
- Required number of iteration until convergence of the retrieval results

A. Dataset

Although a large number of datasets exist for general image and video retrieval, there was no suitable dataset available for multimodal person search and retrieval. Thus the VALID database[4], developed for multimodal biometrics [12], was chosen for the experiments. Although it was developed for a different purpose, it has similar characteristics as in the application scenarios.

It consists of 1060 audiovisual sequences containing individual persons in head and shoulder view either saying a short sentence or counting numbers. Each of the 106 individuals (27 female, 79 male) is captured in 5 environments (1 studio, 4 office), leading to 10 files for each of them. Both the acoustical (noise, reverberation) and visual characteristics (illumination, background) of the environments are quite diverse, making the data even more realistic for the given application scenarios. Figure 3 shows some samples of the same individual within the different environments.

B. Methodology

The main goal of the experiments is to evaluate the performance of the different search and retrieval paradigms integrated into the system and to compare the performance of the different modalities with each other. More specifically three different aspects are considered:

- Quality of ranking relevant documents before non-relevant documents
- Speed of ranking improvement during successive feedback rounds
- Complexity of the interaction depending on the result set size and the number of iterations

[4]http://ee.ucd.ie/validdb/

It is well known, that the policy of the user providing relevance feedback can have a strong impact on the evaluation results [7]. In order to compute reproducible results, an automatic evaluation process without and with user interaction was used. It assumes a stoic user that marks all samples within a result set correctly. Since this is less realistic than just marking some samples and even making mistakes, the obtained results can be seen as an upper bound performance, which may not be achieved in reality. Nevertheless, since the number of relevant items is quite small (10 samples) the real performance will be quite close to this limit. Each sample has been considered for the initial query by example and the mean performance measures have been computed across the resulting 1060 retrieval runs.

The evaluation is based on typical retrieval measures [13]. Both precision/recall and rank based measures have been considered. The most suitable measures are the *average precision* (AP) $\overline{P}$, which measures the average ratio between relevant and retrieved items at the position of relevant items and the *normalized average retrieval rank* (NAR) $\widetilde{R}$, which computes the average rank of relevant items normalized by the number of overall items in the database. Since both measures consider only relevant items and their position within the ranked result list, they can be seen as a combination of recall and precision into a single measure. The former is defined as

$$\overline{P} = \frac{1}{N_R} \sum_{i=1}^{N_R} P_i = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{TP_i}{TP_i + FP_i} \tag{7}$$

with the number of relevant items N_R and the number of true positives TP_i and false positives FP_i at position i. The latter is defined as

$$\widetilde{R} = \frac{1}{NN_R} \left(\sum_{i=1}^{N_R} R_i - \frac{N_R(N_R-1)}{2} \right) \tag{8}$$

with the number of all items N and the rank R_i at position i.

C. Results

Table I provides a comparison of the different retrieval approaches and modalities by showing the average precision and the normalized average retrieval rank after 5 iterations of the relevance feedback. Comparing the different relevance feedback techniques with the query by example paradigm shows a large performance improvement over all modalities. For the audio modality the average precision improves by 22 and 60 % for the SG and the SVM based RF respectively. For the video modality an improvement of 32% for the SG and 49% for the SVM approach are achieved. The gain is even larger for the multimodal system with 43% and 63% for the SG and the SVM respectively. Comparing the different modalities to each other generally shows a performance gain of the multimodal approach with regard to the unimodal approaches (audio, video). While the improvement is only marginal (1.7%) for the query by example it is much larger for the SG (12%) and the SVM (19%) based relevance feedback. The provided results are achieved with the best fusion method for each of the approaches. While the sum rule is the best fusion method for the QBE and SVM approach, the product rule is the most suitable for the SG approach.

Modality	Approach	AP	NAR
Audio	QBE	0.196	0.280
Audio	SG	0.425	0.226
Audio	SVM	0.803	0.089
Video	QBE	0.317	0.126
Video	SG	0.633	0.062
Video	SVM	0.802	0.036
Multimodal	QBE	0.324	0.168
Multimodal	SG	0.753	0.087
Multimodal	SVM	0.991	0.001

TABLE I
COMPARISON OF THE DIFFERENT APPROACHES AND MODALITIES BASED ON THEIR BEST PERFORMANCES.

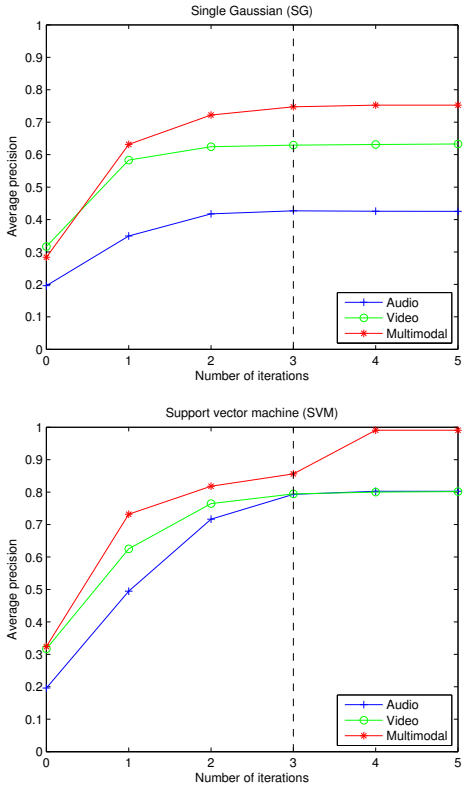

Fig. 4. Comparison of the different approaches and modalities over the number of iterations for a result set size of 45.

Figure 4 provides a more detailed view of the retrieval process by plotting the average precision vs. the number of iterations. Iteration 0 corresponds to the initial query by example and iteration 5 to the final relevance feedback results reported in table I. The results are shown for the maximal result set size of 45. As it can be seen for both the SG and the SVM based approach, the performance converges after 3-4 iterations against a maximum. While the video modality achieves a higher performance than the audio modality for the SG approach, both modalities are comparable for the SVM

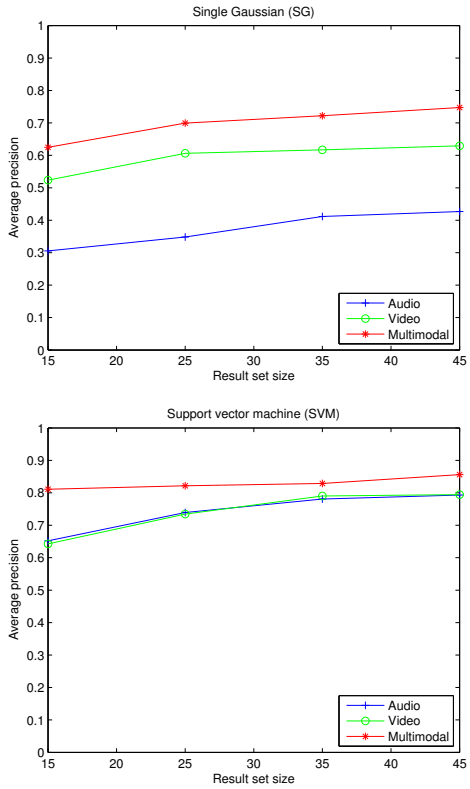

Fig. 5. Comparison of the different approaches and modalities over the result set size for a the 3rd iteration.

based approach.

Figure 5 focuses on another aspect of the retrieval process by plotting the average precision vs. the result set size for the 3rd iteration. As expected the performance increases for larger result set sizes, since more samples are provided as feedback which leads to better models for the user's search intention. For the SG approach the performance of all modalities varies about 10% across the different result set sizes. For the SVM approach the variation of the unimodal systems (15%) is larger as for the multimodal system (5%). Nevertheless it is interesting to see that a larger result set size does not influence the performance too much, which allows to reduce the user's efforts without large performance drops.

Finally, are short analysis of the retrieval efficiency considering the number of RF iterations, the result set size and the required browsing of the samples is provided. With the assumption that it takes the user 4 seconds to play a single sample and judge it to be either relevant or irrelevant, a single iteration for a result set size of 25 items takes about 100 seconds. Considering 3 iterations as the average, a complete search and retrieval session takes about 300 seconds or 5 minutes. In comparison to manually searching through all the 1060 items in the database which takes about 4240 seconds or 70 minutes the effort is reduced by a factor of 14. While the retrieval performance may decrease for larger databases, the complexity reduction will be even larger.

IV. Conclusion

An original system for multimodal person search and retrieval supporting query by example and relevance feedback has been developed. It allows for an efficient search of persons based on their voice and face characteristics. Beside query by example two different relevance feedback techniques (one class and two class) have been integrated and compared to each other. Furthermore, the individual modalities (audio, video) are fused considering different score combination methods.

The experiments show that relevance feedback can improve the retrieval performance considerably with respect to the simple query by example. An even larger improvement is achieved by fusing the individual modalities. Regarding the different RF approaches, the support vector machine constantly outperforms the single Gaussian. In summary the results show that multimodal person search using relevance feedback is feasible and provides a more efficient and reliable way to manage video blogs, podcasts and news broadcasts.

Acknowledgment

The research leading to these results has received funding from the European Communission through the 6th and 7th Framework Programme under grant agreements 038398 (VISNET II) and 216444 (PetaMedia).

References

[1] M. S. Lew, N. Sebe, C. D. Lifl, and R. Jain, "Content based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 1, 2006.

[2] K. W. Bowyer, K. I. Chang, P. Yan, P. J. Flynn, E. Hansley, and S. Sarkar, "Multi modal biometrics: an overview," in *Workshop on Multimodal User Authentication (MMUA)*, 2006.

[3] L. Goldmann, A. Samour, and T. Sikora, "Towards person google: Multimodal person search and retrieval," in *International Conference on Semantics and Digital Media Technologies (SAMT)*, 2007.

[4] P. Delacourt, D. Kryze, and C. J. Wellekens, "Speaker-Based Segmentation For Audio Data Indexing," in *ESCA ETRW Workshop*, 1999.

[5] L. Goldmann, U. J. Mnich, and T. Sikora, "Components and their topology for robust face detection in the presence of partial occlusions," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, 2007, published.

[6] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[7] M. Crucianu, M. Ferecatu, and N. Boujemaa, "Relevance feedback for image retrieval: A short survey," INRIA Rocquencourt, Tech. Rep., 2004.

[8] Z. Su, H. Zhang, S. Li, and S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," *IEEE Transactions on Image Processing*, vol. 12, no. 8, Aug 2003.

[9] G. Guo and S. Z. Li, "Content based audio classification and retrieval by support vector machines," *IEEE Transactions On Neural Networks*, vol. 14, no. 1, Jan 2003.

[10] J. Kittler, H. Mhamad, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.

[11] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, 2005.

[12] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "The realistic multimodal valid database and visual speaker identification comparison experiments," in *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2005.

[13] H. Mueller, W. Mueller, D. McG. Squire, and T. Pun, "Performance evaluation in content based image retrieval: Overview and proposals," University of Geneva, Tech. Rep., 1999.