Multimed Tools Appl (2013) 63:501–520
DOI 10.1007/s11042-011-0834-x

Role-based identity recognition for TV broadcasts

Tobias Schwarze · Thomas Riegel · Seunghan Han · Andreas Hutter · Stefanie Nowak · Sascha Ebel · Christian Petersohn · Patrick Ndjiki-Nya

Published online: 20 July 2011

Abstract Semantic queries involving image understanding aspects require the exploitation of multiple clues, namely the (inter-) relations between objects and events across multiple images, the situational context, and the application context. A prominent example for such queries is the identification of individuals in video sequences. Straightforward face recognition approaches require a model of the persons in question and tend to fail in ill-conditioned environments. Therefore, an alternative approach is to involve contextual conditions of observations in order to determine the role a person plays in the current context. Due to the strong relation between roles, persons and their identities, knowing either often allows inferring about the other. This paper presents a system that implements this approach: First, robust face detection localizes the actors in the video. By clustering similar face instances the relative frequency of their appearance within a sequence is determined. In combination with a coarse textual annotation manually created by the broadcast station's archivist the roles and consequently the identities can be assigned and labeled in the video. Starting with unambiguous assignments and cascading, most of the persons can be identified and labeled successfully. The feasibility and performance of the role-based person identification is demonstrated on the basis of several programs of a popular German TV show, which consists of various elements like interview scenes, games and musical show acts.

Keywords Identity recognition · Metadata · Searching · Clustering · Television programs · Face localization · Labeling

Part of the content of this paper has been presented on 3rd International Workshop at the Automated Information Extraction in Media Production, AIEMPro'10, Florence 25–29 October 2010.

T. Schwarze · T. Riegel (✉) · S. Han · A. Hutter
Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 80200 Munich, Germany
e-mail: Thomas.Riegel@siemens.com

S. Nowak
Fraunhofer Institute for Digital Media Technology, Ehrenbergstrasse 31, 98693 Ilmenau, Germany

S. Ebel · C. Petersohn · P. Ndjiki-Nya
Fraunhofer Institute for Telecommunications, Einsteinufer 37, 10587 Berlin, Germany

1 Introduction

Massively growing amounts of rich multimedia data available in broadcast archives and video portals increase the need for sophisticated indexing mechanisms for content search and retrieval scenarios. Available systems lack an understanding of the data and thus stay decoupled from the actual video content.

Today, when content-based queries are posed to multimedia archives, they are answered based on manual user descriptions, mostly in the form of textual annotations like keyword labels or captions that are attached to the content. Manual annotations are usually neither complete nor accurate and often affected by the subjective view of the annotator. Even more prohibitive is the fact that the manual annotation of video footage is a very time consuming and thus costly task. Annotating one hour of footage costs an annotator around four working hours-reason enough to automate the process as far as possible.

The level of annotation that can be achieved automatically by dedicated visual analysis algorithms, however, is mostly not sufficient to manage video content in a satisfying way. The derivation of semantic data is required to enable a natural way of interacting with the retrieval system. It means that the systems must be able to process queries which express what the user is looking for, not how the objects in question look. This leads to the well known problem of bridging the semantic gap between formal, machine understandable concept representations and human real world concepts.

In this paper, we address the domain of TV shows and particularly the problem of identifying the participating persons in order to label them with their names, which in return allows the retrieval of content related to a specific person. We follow a strategy of exploiting and combining all kinds of available information, from visual data analysis to the integration of related sources as given by context and domain background knowledge. The paper is organized as follows: Section 2 gives a short review of the state of the art in semantic video understanding approaches. Section 3 introduces a general system architecture for video search and retrieval applications, before we detail our reasoning mechanism in Section 4 and its implementation in Section 5. The results are evaluated in Section 6 and discussed in Section 7.

2 Related work

In general, the approaches towards semantic video understanding like identifying persons can be categorized into two classes.

Bottom-up approaches in video analysis use the low-level metadata indexes from temporal (usually shot detection) and regional (object detection) decomposition of one or more modalities in order to infer semantic metadata. Usually prior knowledge about the reflection of semantics on pixel level is used to design special purpose classifiers. Such a classifier may, e.g., support the classification into genres. For commercial detection, classifiers based on the average shot length, edge change ratio and motion vector length have been used to identify the high dynamics usually found in commercials in [20]. Slow motion passages and the amount of textual overlays, together with a reporter's voice, may be hints for sport broadcasts as described in [19]. Javed et al. [16] exploit the highly repetitive structure of talk shows to separate commercials and find shots showing the show host.

Transferring the bottom-up strategy to the problem of identifying persons in a TV show leads to the classical domain of face recognition. Starting off with a detected face, the

system tries to match it against a set of known individuals to evaluate the identity. This direct approach is pursued by Houghton in his Named Faces system [14], which is building a database of named faces by recognizing the people names overlaid on the video frames using video optical character recognition (VOCR). Similarly Boujemaa [5] et al. realized a hybrid thesaurus (textual and visual) approach based on face detection and recognition to provide archivists online central and updated references for most frequently encountered humans in video news. Their accuracy stands and falls with the reliability of the employed face recognition approach. Face recognition has been a major research topic for a long time; the range of approaches is very scattered accordingly. The accuracy of many face recognition approaches in constrained domains is reported reasonable, for example [2, 3, 10, 13, 18], but as soon as the capturing conditions become more varying (e.g., lateral/elevated viewing angle, bad lighting conditions, occlusions, etc.) the recognition rates decrease drastically. In recent surveys of face recognition techniques [7, 33], especially pose variation, which occurs in TV casts, was identified as one of the prominent unsolved problems in the research of face recognition. Bearing that in mind and as our approach copes without a facial recognition component, a review of different pose-invariant face recognition techniques is beyond the scope of this paper. Nevertheless, for further reading a recent and well elaborated survey especially on pose-invariant face recognition approaches and according references can be found in [32].

One way to tackle the problem of face-name association is to exploit the relations between videos or images and the associated texts in order to label the faces with names under less or even no manual intervention. Name-it [27] is the first proposal on face-name association in news videos based on the co-occurrence between the detected faces and names extracted from the video transcript. A face is labeled with the name which frequently co-occurs with it. Yang et al. [29] employed the closed caption and speech transcript to build models for predicting the probability that a name in the text matches to a face on the video frame. By using multiple-instance learning for partial labeled faces, the effort of collecting data by users can be reduced. In [24], the speech transcript was also used to find people frequently appearing in the news videos. Similarly, for face identification in news images, the problem was also addressed as clustering or classifying the faces to the people's specific appearance models, supervised by the name cues extracted from the image captions or associated news articles [4, 11, 15].

Hence, many efforts on film analysis were devoted to the detection of main characters, automatic cast listing, movie segmentation, or summarizing of a feature film as a video abstract but not assigning real names to them. Arandjelovic and Zisserman [1] used face images as a query to retrieve particular characters. Affine warping and illumination correction were utilized to alleviate the effects of pose and illumination variations. To tackle the problem of story segmentation, Chaisorn et al. [6] proposed a two-level multi-modal approach by analyzing first the video at the shot level using a variety of low- and high-level features, and classifying the shots into pre-defined categories using a Decision Tree. Subsequently the news story boundaries have been identified by performing HMM. Lehane et al. [21] went one step further and searched for specific scenes. They described an approach for detecting dialogue scenes in movies using automatically extracted low- and mid-level visual features that characterize the visual content of individual shots, and which are then combined using a state transition machine that models the shot-level temporal characteristics of the scene under investigation. Lienhart et al. [22] tried to detect and classify special events such as dialogs, shots, explosions and text of the title sequence for the automatic production of a video trailer of a feature film. First the input video is segmented into larger semantic units, so-called shot clusters or scenes, followed by the detection and extraction of contained semantically rich special events.

Top-down reasoning approaches are inspired by the fact that the surrounding conditions of an observation are of major importance when trying to understand and interpret its meaning. Each object appears in a context that is given by its surrounding environment and the interaction and relation to other objects.

Reasoning in a top-down manner requires the incorporation of high-level knowledge about the domain. Predominantly this knowledge must capture the entities that make up the components of the world, their interrelations and interactions.

A context model comprises this domain knowledge. The general goal is to find an assignment between detected object concepts or segmented image regions, and the semantics of the context, so that the model is matched. Scenario recognition could build upon this by specifying a set of roles which are taken by a set of involved persons and which are characterized by a specific behavior or other properties like probabilities or their frequency of appearance.

In order to evaluate such an assignment, usually more knowledge about the objects has to be gathered. This results mostly in a bottom-up analysis step to extract additional features and object characteristics. External sources of information can be a valuable support. Subtitles and transcripts contain high-level semantic data which can facilitate identification and labeling tasks, as seen in [8]. Different from most state-of-the-art methods on naming faces in the videos, which use the local matching between a visible face and one of the names extracted from the temporally local video transcript, Zhang et al. [12] attempt to do a global matching between names and clustered face tracks in case that there are not enough local name cues. A graph matching method has been utilized to build name-face associations between the name affinity network and the face affinity network which are, respectively, derived from their own domains (script and video). A comparison with the aforementioned local matching approach from Everingham [8] has been carried out by the authors, showing that the overall performance is comparable while requiring less information (film script) than the others (audio transcript + subtitle). Further on they report at the high levels of recall the preeminence of their method, due to the effectiveness of the face track distance measure in clustering and the employment of the multi-modal features in cluster pruning. Both approaches depend on considerable image analysis and quite detailed textual annotation (film script/transcript + subtitle), which typically is not available for TV broadcasts.

Rather than focusing on extensive data analysis in a bottom-up manner, our approach towards identity recognition focuses on the integration of surrounding context knowledge in order to reason about the function the persons play during the show. Inferring the role of a person eventually allows us to identify the person in a modeled context. Before describing these reasoning mechanisms in detail in Chapter 4, in Chapter 3 we briefly introduce the video search and retrieval system we use.

3 Video search and retrieval system

We propose to employ a system architecture for video search and retrieval (Fig. 1) that was derived from the generic architecture introduced in [30] for surveillance applications. The components of the system are briefly described in the following.

The starting point is always the acquisition of the video footage and the generation and storage of metadata describing the video. The video is either captured by a number of cameras (e.g., in surveillance applications) or is available as a single, readily composed stream in case of TV broadcasts. In either case, the footage is stored in a *Video Archive*. The

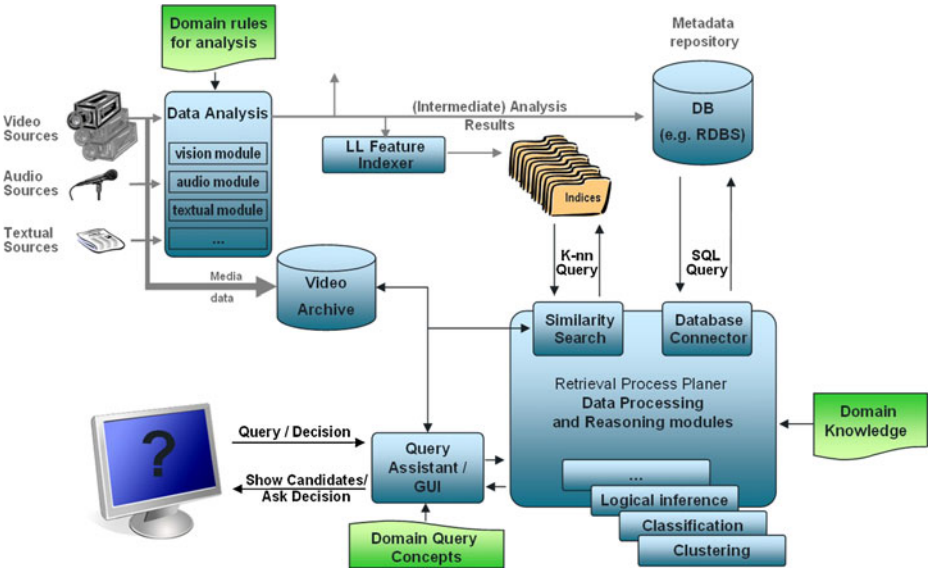

Fig. 1 Search and retrieval system architecture

two cases differ in that surveillance footage is usually processed on the fly, while a broadcast stream is mostly processed after production.

A set of *data analysis* modules processes the video data in order to extract basic metadata instances. Their semantic level can range from low signal-based features like color descriptors, textures descriptors or shot boundaries [25, 30], up to object concepts like "cars" or "faces", or scene concepts like "audience". Which modules have to be applied depends on the domain and is therefore supplied by external rules. The set of analysis modules generates the majority of metadata that is necessary for later post-processing and reasoning. The analysis is not limited to the visual domain. Audio and text analysis modules like optical character recognition can be applied also.

The metadata are stored in the *metadata repository.* There are multiple ways to store metadata; the most common ones are relational databases or, increasingly, RDF triple stores. The metadata have to be synchronized with each other and especially with the associated video data, e.g., by attaching the media time in form of a global time stamp to the metadata instances. The storage facilitates accessing the metadata by querying the storage with well-known query languages like SQL, or SPARQL in case of RDF storage solutions. Furthermore, it allows combining the different kinds of metadata to form new metadata instances; it allows updating existing instances, respectively inserting new instances and narrowing down search spaces by powerful selection and sorting mechanisms.

Implicitly connected to the storage is the *indexing mechanism* that supports the important task of finding items matching the query pattern, but in the case of visual features also items similar to a given item.

Selecting particular instances from the metadata repository by specifying ranges of media time, object size, object locations or other low-level indexes, and selecting the top ranked items from a similarity search with a given sample form a powerful foundation for further reasoning activities. The core part of the system architecture, therefore, consists of a module that steers the process to answer a query posed to the system. It coordinates the

retrieval of appropriate metadata from the repository and the data post-processing by passing the metadata to a number of data processing and reasoning modules. The output of these processes is a ranked list of candidates that match the posed query and can be presented to the user. The reasoning techniques used here rely on the knowledge represented by context and domain models.

A query composition module eases the interaction with the system. Since the task of understanding the user's intention of a query posed in natural language leads back to the problem of bridging the semantic gap between user input and its meaning, the system assists the user in posing the query in a machine understandable way. The query composition could be eased by restricting the search to certain keywords or by offering a list of suggested entities, as for instance all appearing guests in a TV show.

The following chapters specifically address the post-processing and reasoning part of this architecture.

4 Identity reasoning

Every person that participates in a show takes on a roll out of a set of known roles, which are given by the show concept. Thus there exists a strong relation between roles, persons and their identities. Knowing one often allows inferring about the other. It is important to distinguish between persons and identities at this point. From a video analytical point of view, a person is an object that can be detected by some known properties like its shape, or the presence of a face. The identity in contrast refers to a specific person that may disappear and reappear during the show. Hence a person is identified via its identity.

Roles and identities are forms of metadata that visual analysis cannot obtain by plain observation. Knowledge about the role and/or identity of an object is metadata on a high semantic level and a big step towards understanding the content.

The classical approach towards labeling persons with their identity is based on the assumption that the identification can be achieved by comparing their face or some other descriptive features with a database of known identities. In case the person is unknown to the system the identification fails. On the other hand, in a top-down model based approach, the context in which a person was detected should be considered. The context of a broadcast TV show prescribes a set of roles that are taken by the identities which are participating in the show. If the mapping between roles and identities is known or can be gathered from external sources, the identity labeling can be largely achieved by estimating the role of a person under observation. Therefore, the key task is not to identify the person itself, but rather to infer the person's role, which can then be used for identification in the given context.

The presence and characteristics of the roles can be modeled in a context model. Usually it will be necessary to gather additional knowledge about the observed persons in order to infer the role they play.

In our given domain of TV shows we detect persons by the appearance of faces. A face detection module is used to build an index of all faces; details are further described in Section 5.1. The single face detections do not contain much information on the role the corresponding persons play in the given context. Single persons are represented by many independently detected faces. A first processing step should accordingly map the detected faces to classes of persons whose roles and identities are still unknown. This enables to further investigate the characteristics of the unknown identities, in order to infer their role and eventually infer their identity.

The complete workflow is shown in Fig. 2 and further detailed in the following.

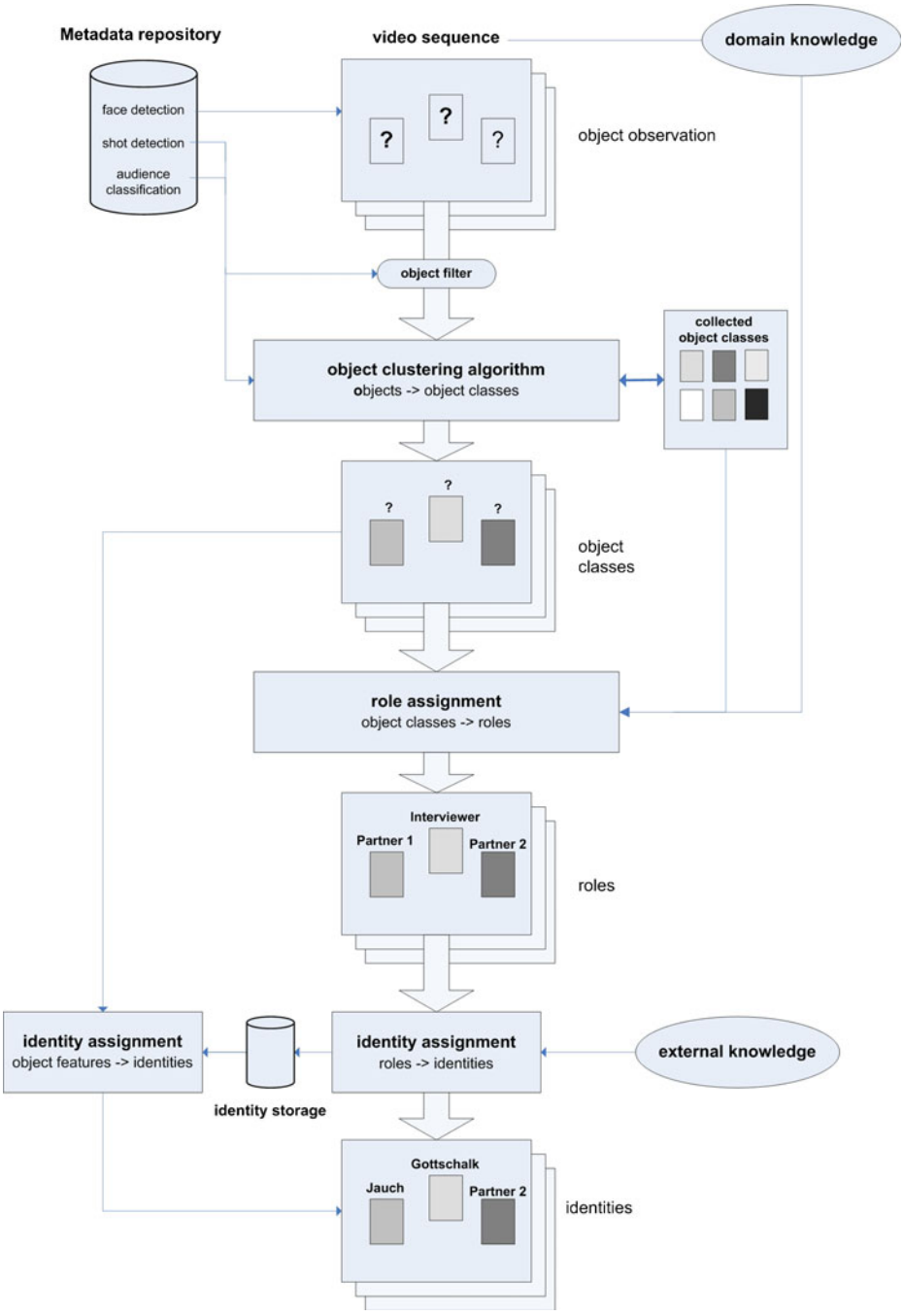

Fig. 2 Workflow for role-based identity recognition

4.1 Face clustering

The task of mapping detected faces to unknown—i.e., unlabelled—identities is equivalent to grouping the faces into classes that each show a single identity. A clustering algorithm yields such an output. It requires a similarity measurement to classify two faces as equal or not. This step can be performed by matching visual properties in combination with other clues and constraints from the context. Details on the implementation of clustering and similarity measurement are given in Sections 5.2 and 5.3.

The result of this step is a varying number of clusters, each of them containing face instances that show a distinct person, along with a face model that represents the person by means of a feature vector.

These clusters facilitate two subsequent steps: A classical similarity lookup by matching the face models against a database with already known identities can be used in order to assign an identity label to a cluster; or, as further described here, the role the represented person plays can be inferred.

4.2 Role assignment

The characteristics employed to infer the role naturally depend on the kind of context. As we are dealing with TV shows, we can make use of the characteristics of interview scenes. An interview context typically contains two sorts of roles: an interviewer, and at least one interviewee, or guest. The important and common property of interview scenes is the strong focus on the interviewee, which leads to clearly more frequent occurrences of the interviewed person in the video. This can be explained by two facts:

1. The viewer of an interview broadcast is typically interested in the interviewed person, not in the interviewer. This results in a preference of shots showing the interviewee.
2. Answering a question usually takes much longer than posing the question. This results in longer and more frequent appearances of the interviewee.

In case of an interview between two persons, this statistical knowledge has been found robust to identify both roles, without having any additional knowledge about the persons themselves. When applied to the clustering result of interview scenes, this corresponds to analyzing the clusters' sizes in terms of the total number of faces included. The biggest cluster can then be assigned to the role "guest" or "interviewee". In case of only two persons the remaining next biggest cluster will probably hold the interviewer. In case there are more than two people involved, the role assignment becomes ambiguous here and requires considering additional contextual information.

4.3 Identity assignment

As previously mentioned, the final identity assignment is performed once the role of a cluster is known. In case of persons in a TV show the obvious identity label would be the name of the person. Usually, additional information about the show is available in program magazines, reviews, supplemental program information or other sources which enable the extraction of names of show host, guests, and in many cases also the order of their appearance in the show.

The role assignment allows identifying the guest in all of the interview scenes. Depending on the number of other guests participating passively in the interview, also the

show host can be identified by his frequency of appearance, but usually there is ambiguousness between the interviewer role and other guest roles. To solve this issue, we use an *identity storage* that stores all persons which have been already identified by their guest role. Clusters which cannot clearly be assigned to a role are matched against this storage to obtain their identity.

Thus the identity labeling is treated from two sides: A primary labeling step based on the inferred role of the person, which is obtained from the statistics of the clustering result, and an alternative labeling step based on a visual comparison with previously identified and labeled persons.

In consequence this means that the processing does not solely depend on absolute completeness of the clustering outcome. A weak clustering due to heavy changes in illumination or pose within a scene might result in persons being scattered among two or three clusters instead of one. By matching all of these clusters against the identity storage the correct label can often still be assigned to all of them, if the person was identified before and added to the identity storage. This allows the clustering similarity threshold to be chosen more strictly in order to avoid mixed clusters with faulty assignments.

Furthermore there can be multiple representations for each identity in the identity storage to consider possible changes in their appearance during the program.

This proceeding follows the idea of exploiting all different kinds of clues rather than relying on a single source of knowledge. Both steps taken together provide the means to automatically annotate large parts of the show with high level semantic indexes.

5 Video analysis

The proposed workflow requires some basic techniques that need to be adapted to the scenario of persons in a TV show.

A basic requirement is the reliable extraction of the objects of interest, i.e., persons in this case. Face detection and localization algorithms have reached a sophisticated reliability level and enable the detection of faces even in varying poses. Thus, we utilize a face localization module to build a person index, which forms the main part of our metadata foundation for further reasoning steps.

The second important part is the clustering algorithm, which builds the input for the interview-model based role and identity labeling.

5.1 Face detection and localization

The localization of faces in videos is based on a face detector that works on single images. The face detector uses sliding windows of different sizes to generate regions of interest that are given to a classifier. The classifier then decides if the given region of interest represents a face. It uses a cascade structure similar to that proposed by Viola and Jones [28], but in contrast to them our detector performs a multi-class classification. That is, faces in multiple poses can be located (left profile, left oblique, frontal, right oblique, right profile). The detector was trained using a special boosting technique called MBHboost [23]. In this scenario LBP-features [31] were utilized, because they yielded better results than the Haar features originally used by Viola and Jones.

The resulting metadata are the position of the detected face along with the size of the estimated bounding box and an ID. As camera moves are usually smooth, it is assumed that size and location of a face do not change significantly from one frame to the next. We use

this information for a simple tracking approach: A full processing takes place every eighth frame only. For the remaining frames only those regions surrounding the locations of faces of the previous frame are searched. This way, a detected face that is recognized again in subsequent frames of a shot can be treated as a single detection and will be assigned the same ID. However, new IDs are assigned with the start of each new shot or when the tracking mechanism fails. Shot boundaries are identified employing the temporal video structure detection system described in [25].

5.2 Clustering

We implemented the clustering based on a similarity measure between persons and logical facts that constrain the cluster assignment.

First and foremost, it can be stated that two detected faces which appear simultaneously cannot belong to the same identity—assuming that there is no mirroring or background screens present in the scene. The rare occurrence of these cases leads to an additional cluster containing the mirrored person which is then usually not considered in the role assignment due to its small size.

Furthermore it can be assumed that no person of interest is located within the audience. This justifies excluding shots showing more than a predefined number of faces from the clustering in order to avoid faulty assignments of random audience members to one of the persons of interest. These constraints are taken into account to support a visual similarity measure, which gives the strongest clue here.

The clustering algorithm takes the output of the face localization module and processes the detected faces in their chronological order of appearance. The feature descriptor described in Chapter 5.3 is extracted from the video for every detected face in order to compare it with the existing clusters. Thereby the simultaneous appearance constraint is considered, which possibly excludes some of the clusters. Based on the distance values a ranking is created to find the cluster that matches the face best. If the distance to the best matching cluster is below a predefined threshold, the face is assigned to this cluster. Otherwise a new cluster is created with the face as first member. Every cluster maintains a face model comprising the mean feature descriptor of all added faces, which also allows comparing clusters among each other. After adding new faces to a cluster, a merging operation is performed to fuse clusters that converged thereby.

5.3 Similarity measurement

A visual similarity measure for persons mainly requires invariance against changes in size and pose. Though specific approaches for face features have been developed specifically for recognition tasks, additional facts should be considered as well. Contrary to the classic identification approaches, we do not aim at recognizing the person, but instead verifying the person as being identical to one of the persons seen before. Other than previously seen approaches (e.g., [9]) the clustering is not intended to be applied on the whole show. The main aim is to reason about the role of persons in a known context as given here by interview scenarios. Thus also features can be considered which are representative only during the particular program, like the color of hair, or the dress the person wears.

Based on the face localization, we extract a hair and a chest region from the image to compare two persons, as shown in Fig. 3. A feature vector for both regions was calculated and constitutes an extremely simplified person model.

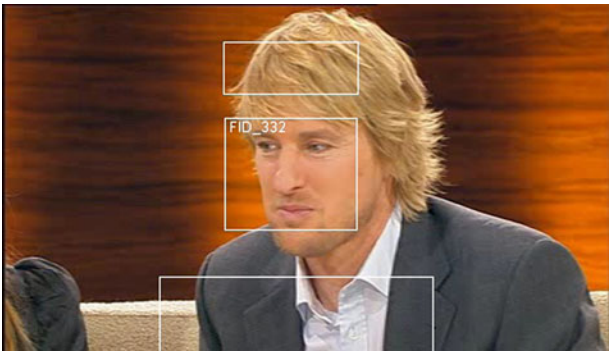

Fig. 3 Detected face region and associated hair and chest region for similarity estimation

The feature descriptor must exhibit two functionalities in order to be applicable for the clustering problem: First, a distance metric must exist which defines how two descriptors are to be compared. Second, it must be possible to calculate an intrinsic mean of multiple descriptor instances in order to build a model that represents the objects contained in a cluster.

The feature descriptor recently proposed by Tuzel, Porikli and Meer [26] fulfils these requirements. They propose to use the covariance of several image statistics computed inside a region of interest as region descriptor. The covariance matrix provides an elegant and natural way to fuse different image features and represent a region with a low dimensional descriptor. In general, a single covariance matrix is stated to be sufficient to represent a region in different views and poses. Regions of different size can also be compared since the dimension of the covariance matrices only depends on the number of features used, which makes them scale invariant.

The covariance matrix is calculated from a $W \times H \times d$-dimensional feature image comprising any d extracted image features, like, e.g., color layers, derivatives, edge magnitudes or orientations. In our implementation, a $d=5$ dimensional feature vector f is constructed for every region comprising the hue, saturation and intensity value as well as its pixel coordinates. Adding the pixel coordinate values as a feature vector allows considering the spatial arrangement of image features.

$$f_k = [x, y, H(x,y), S(x,y), V(x,y)], \quad k = 1 \ldots WH. \tag{1}$$

The covariance matrix C of dimension $d \times d$ for a rectangular region of size $W \times H$ is then calculated by

$$C = \frac{1}{WH} \sum_{k=1}^{WH} (f_k - \mu)(f_k - \mu)^T, \tag{2}$$

where μ is the vector of feature means for all pixels in the image region. This results in a symmetric matrix with the diagonal entries representing the variance of the feature vectors and the non-diagonal entries the correlation between the corresponding features.

The distance measure for two regions i and j uses the sum of squared logarithms of the generalized eigenvalues as

$$\rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)}. \tag{3}$$

with the d generalized eigenvalues λ_k of C_i and C_j calculated from

$$\lambda_k C_i x_k - C_j x_k = 0, \quad k = 1 \ldots d. \tag{4}$$

In order to obtain a "face" model from covariance matrices a procedure is needed which determines an average covariance descriptor from a set of covariance matrices. An algorithm for this purpose that iteratively approximates the intrinsic mean of multiple covariance matrices is also described in [26].

6 Experimental results

We evaluate the proposed identity reasoning algorithm with seven episodes of a popular German TV game show. The show includes various elements like interview scenes, games and musical show acts. The footage sums up to around 18 hours. In total, around 40 celebrity guests appear. The above mentioned face localization module is used to detect and store the generated metadata in a first pass, roughly resulting in 47,500 face IDs which belong to the 1,400,000 detected bounding box instances. A coarse textual annotation (script) written by the broadcast station's archivist is used to identify the interview scenes and obtain the names of the participating guests. Overall, it is possible to label almost all of the appearing guests with their corresponding name based on this midlevel metadata foundation and by applying the proposed identity reasoning chain.

A direct comparison with an existing approach would be desirable. However this will be a rather delicate and difficult issue, because corresponding approaches are adapted to their application domains inherently. Many selected the news scenario due to the advantageous rigid structure (e.g., [4, 5, 24, 29]) and more advanced approaches (e.g., [8, 12]) assess featured film showing detailed script data. Unfortunately, no comparable approaches dealing with game shows, which feature a less rigid structure and only coarse script data, could be identified. Tailoring other approaches accordingly would be-beside the implementation effort-like comparing apples to oranges. Nevertheless a comparison shall/will be part of future work.

The next sub-chapters give detailed results on the overall performance. For evaluation, we distinguish between the two steps role-assignment and identity-labeling.

6.1 Evaluation of the role assignment

The most important step, the role assignment, depends on the accuracy of the clustering algorithm which employs the similarity measure described in Section 5.3. The evaluation uses as test corpus one episode of "Wetten dass...?" which was captured in Düsseldorf on the 28th of February, 2009 and is manually assessed. Every face that was detected by the face detection module is assigned to a label of one of the celebrities taking part in the show. As just shots are considered that were retrieved by the face detection component, it is assured that the evaluation considers the error that is made in clustering and later stages of the program workflow. Errors from missed faces in the face detection are not accumulated. In total, the video contains 237,602 frames. 110,113 frames with in total 4,153 face IDs were manually assessed with 10 different celebrities of interest. These 10 persons are depicted at 75,664 frames in 1,034 shots and are referenced by 1,765 face IDs. Because the biggest resulting cluster is always considered to be the searched celebrity, in all other clusters missing faces are counted.

Table 1 shows the results for the cluster evaluation in terms of true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), precision and recall. In the following the results are illustrated in three examples:

Example 1: Search for the person "GOTTSCHALK"

Due to the information in the script, only the beginning and the end of the show is taken into account for the search. The result consists of 16 clusters, the largest having 603 faces, the 2nd 179 faces, the 3rd having 154 faces and so on. The biggest cluster groups only frames with faces of "GOTTSCHALK", but clusters 9, 10 and 12 are false negatives having 53 faces all together. The precision is 100% and the recall is 91.9%.

Example 2: Search for "ANISTON" followed by a search for "WILSON"

The result consists of 15 clusters. The largest one contains 1452 faces of ANISTON, the 2nd 574 faces of WILSON who is being interviewed together with ANISTON, the 3rd cluster contains 476 faces of WILSON and the 4th cluster contains 262 faces of the interviewer GOTTSCHALK. All 15 clusters contain only faces of one person per cluster. The precision is 100%. But 166 faces of ANISTON are clustered in other clusters, which lower the recall to 89.7%.

The search for "WILSON" is performed on the same time span with the same clustering result as the search for the person "ANISTON". As the biggest cluster contains solely faces of the person "ANISTON", precision and recall have values of 0% for "WILSON".

Summarizing, the clustering approach has a very high precision, with 100% for seven out of ten persons and an average of 79.5%, while the recall ranges between 0% and 91.9% with a mean of 46.4%. In two cases, the assumption that the biggest cluster contains the person which was searched for is violated and results in a precision and recall of 0%. This is typical for joint interviews with two or more guests and will be tackled in future work.

6.2 Evaluation of the identity-labeling

In joint interviews with an unknown number of guests, ambiguousness arises between the interviewer and other guest roles. This ambiguousness means that the clustering alone does

Table 1 Results for the cluster evaluation

Celebrity Name	# clusters	TP	FP	FN	TN	Recall	Precision
ANISTON	15	1452	0	166	1985	89.7%	100.0%
WILSON	15	0	1452	1511	640	0.0%	0.0%
GOTTSCHALK	16	603	0	53	760	91.9%	100.0%
BECKER	37	2011	0	786	2701	71.9%	100.0%
MAKATSCH	21	2179	123	1179	654	64.9%	94.7%
FERCH	23	675	0	1359	2066	33.2%	100.0%
SAWATZKI	23	0	675	1527	1898	0.0%	0.0%
FREYDANK	21	952	0	1470	1186	39.3%	100.0%
DUFFY	13	317	0	719	47	30.6%	100.0%
BLUM	13	105	0	145	170	42.0%	100.0%

usually not allow assigning the interviewer role (cf. Example 2 above). As additional contextual knowledge, we use the information that the host always appears first during the shows' opening sequences. This fact is exploited in order to find a face model that represents the show host, who then takes the interviewer role in all the interviews.

Once the roles are assigned to the clusters, each person (i.e., cluster) can be labeled with the identity which takes on the role in the current context. In the experiments, our source of external information was the archivist's summarizing annotation script, which contained all names of the active guests, but often also names of passive third persons. Accordingly, the mapping between names and roles is often a matter of text interpretation and pattern matching and can be a source of errors.

An objective quantitative evaluation of the identity-labeling emerges to be very hard due to the few identities appearing in the video data, the simplified assumption that the guest role coincides with the highest appearance frequency, and the error-prone similarity measure described in Section 5.3, which is used to assign identities via the identity storage. Therefore, this test is only performed using a subjective impression by exemplarily testing how the identification quality progresses with a growing person database. Figure 4 depicts a screenshot of the experimental system executing the role-based identity recognition.

The test scenario is explained step by step starting with an empty identity storage.

- We start by searching for the face of "GOTTSCHALK" and approve the name of the first cluster which contains faces of "GOTTSCHALK".
- Restarting the search for "GOTTSCHALK", the experimental system recommends three more clusters with faces of "GOTTSCHALK" to be labeled "GOTTSCHALK", which we approve. Here the similarity matching of the visual features works and suggests more clusters that contain "GOTTSCHALK".
- Subsequently we search for an interview, where "GOTTSCHALK" is the interviewer, for example the interview of ANISTON and WILSON, by searching for "ANISTON". The system proposes the largest cluster to contain faces of "ANISTON", but it also proposes the 2nd largest cluster, which actually contains faces of "WILSON" to be labeled "GOTTSCHALK". Correcting the error by relabeling the second cluster with the tag "WILSON" we restart the search for "ANISTON".
- Now, the third cluster which contains faces of WILSON is proposed to be labeled "WILSON" but cluster 4 with faces of GOTTSCHALK is still not correctly recognized. We further annotate all clusters in the result of "ANISTON" with their correct labels.
- Now we search for the interview with "FREYDANK" which is the next part of the show. The largest cluster which contains faces of FREYDANK is named correctly, but cluster 4 which also contains faces of FREYDANK is assumed to be ANISTON. Faces of GOTTSCHALK in the 6th cluster are recognized correctly, faces of FREYDANK in the 7th cluster are assumed to be "WILSON" and faces of GOTTSCHALK in the 9th cluster are proposed to be labeled "WILSON".
- Labeling and relabeling all clusters correctly, we move on and search for the interview with "BECKER". This time faces of GOTTSCHALK, WILSON, ANISTON and FREYDANK are correctly labeled, but on the other hand some faces of BECKER are assumed to be WILSON.

This test scenario shows that the experimental system works better, the more face clusters are labeled with the correct tags. Faces first labeled are correctly recognized later on. New faces cause errors because they are not contained in the identity storage and no

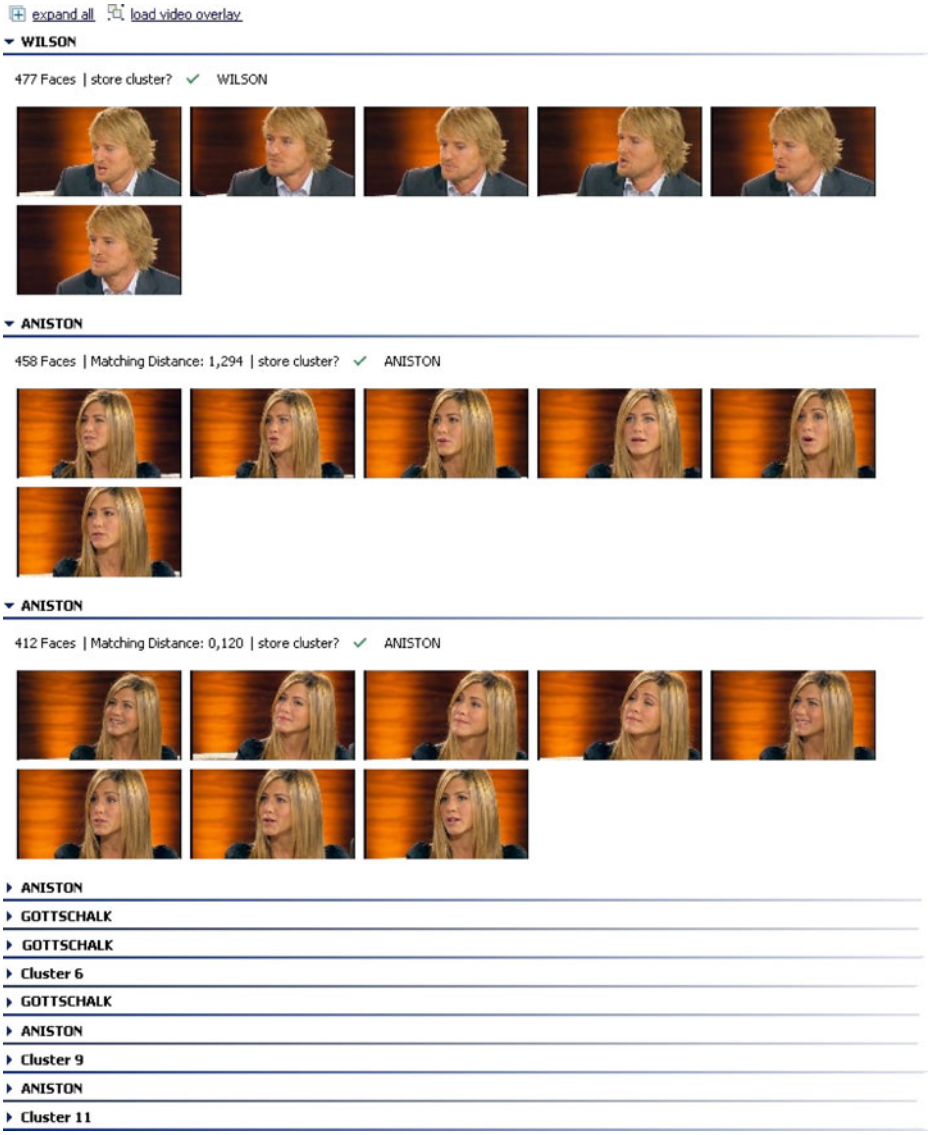

Fig. 4 Screenshot identifying WILSON with already "learned" ANISTON and GOTTSCHALK

inferring rules can be applied. In some cases the similarity matching of visual features helps to predict the correct labels. Sometimes, already stored person labels like "GOTTSCHALK" cannot be identified in other clusters.

At the bottom line the rate of correct identification in joint interviews with two or more guests is about 70%. In the cases where some of the guests have already been successfully identified and labeled in prior scenes, the identity store allows their recognition based on a similarity lookup. By applying this lookup in a second pass, most of the guests' identities in interviews with three or more persons can be identified despite the name vs. role ambiguities.

All in all, a great majority (more than 80%) of the active guests and more than 60% of the passive guests could automatically be labeled with a high completeness within interview scenes. The completeness drops when instances of the same person are grouped into multiple clusters. This occasionally happens when persons are captured with their head turned heavily, which affects the visual similarity measure. In this case only a part of the person's instances is labeled while the instances in the other clusters containing the same person remain unlabeled. Usually these clusters are rather small and consist of only a few "ill conditioned" instances (3–4 compared to 20–30 in a sound cluster).

7 Conclusions and future work

In this paper, we presented a system for automatically analyzing and annotating TV programs by reasoning intelligently upon a closed, well known and modeled domain, namely TV talk shows.

Based on a set of automatically obtainable metadata a post-processing step facilitates the creation of new metadata on a high level of semantics in form of roles and identities of persons appearing in the show.

By leveraging coarse but high-level semantic information, in this case an available manually written summarizing script, the developed system manages to bridge the semantic gap between a detected generic face and the identity it belongs to. The assumptions modeled are derived from a general model of interview situations and are thus applicable to various other broadcast programs as well.

Obviously future work should include the facial image analysis improving particularly the face track estimation (e.g., as proposed in [30]), especially as we focused on the exploration of higher level semantic allowing us to employ straightforward image analysis and to compensate low-level image analysis deficits.

A far more interesting point will be the question, how more general roles can be modeled/represented to query upon, involving the inherent uncertainty of image analysis and the incomplete knowledge representing possible situational/contextual information. We intend to explore logic programming to represent human knowledge and the use of subjective logic [17] to handle uncertainty implied in the extracted data and also of the modeled knowledge itself. And-by doing this-the system shall become flexible enough to be able to compare it with related work realistically.

Acknowledgments This work has been supported by the THESEUS Program, which is funded by the German Federal Ministry of Economics and Technology. In particular, we thank our THESEUS project partner Institut für Rundfunktechnik for providing the TV program data and permission to use them for scientific purposes.

References

1. Arandjelovic O, Zisserman A (2005) "Automatic face recognition for film character retrieval in feature-length films". In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 860–867
2. Bartlett MS, Movellan JR, Sejnowski TJ (2002) Face recognition by independent component analysis. IEEE Trans Neural Network 13(6):1450–1464
3. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720

4. Berg T, Berg A, Edwards J, Maire M, White R, Teh Y, Miller E, Foryth D (2004) "Names and faces in the news". In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, Washington, DC, USA, vol. 2, pp. 848–854
5. Boujemaa N, Fleuret F, Gouet V, Sahbi H (2004) "Automatic textual annotation of video news based on semantic visual object extraction". In: Proc. SPIE Storage and Retrieval Methods and Applications for Multimedia, San Jose, California, pp. 329–339
6. Chaisorn L, Koh C, Zhao Y, Xu H, Chua T-S, Qi T (2003) "Two- level multi-modal framework for news story segmentation of large video corpus". In: Proc. 12th Text Retrieval Conference, Gaithersburg, MD, USA
7. Chen S, Tan X, Zhou Z-H, Zhang F (2006) Face recognition from a single image per person: a survey. IEEE Pattern Recogn 39(9):1725–1745
8. Everingham M, Sivic J, Zisserman A. "Hello! My name is… Buffy—automatic naming of characters in TV video". In: Proc. British Machine Vision Conference, Sept. 2006, Edinburgh
9. Fitzgibbon AW, Zisserman A (2002) "On affine invariant clustering and automatic cast listing in movies". In: Proc. 7th European Conference on Computer Vision, Copenhagen, pp. 304–320
10. Gao Y, Leung MKH (2002) Face recognition using line edge map. IEEE Trans Pattern Anal Mach Intell 24(6):764–779
11. Guillaumin M, Mensink T, Verbeek J, Schmid C (2008) "Automatic face naming with caption-based supervision". In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, Anchorage, AK, USA, pp. 1–8
12. Han S, Hutter A, Stechele W (2009) "Toward contextual forensic retrieval for visual surveillance: challenges and an architectural approach". In: Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, London, United Kingdom, pp. 201–204
13. He X, Yan S, Hu Y, Niyogi P, Zhang H-J (2005) Face recognition using Laplacian faces. IEEE Trans Pattern Anal Mach Intell 27(3):328–340
14. Houghton R (1999) Named faces: putting names to faces. IEEE Intell Syst 14(5):45–50
15. Jain V, Learned-Miller E, McCallum A (2007) "People-LDA: anchoring topics to people using face recognition". In: Proc. IEEE Int. Conf. Computer Vision, Rio de Janeiro, pp. 1–8
16. Javed O, Rasheed Z, Shah M (2001) "A framework for segmentation of talk & game shows". In: Proc. Int. Conf. on Computer Vision, Vancouver, BC, Canada, pp. 532–537
17. Jøsang A (2001) "A logic for uncertain probabilities," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 9(3): 279–311
18. Kirby M, Sirovich L (1990) Application of the Karhunen–Loève procedure for the characterization of human face. IEEE Trans Pattern Anal Mach Intell 12(1):103–108
19. Kobla V, Dementhon D, Doermann D (2000) "Identifying sports videos using replay, text, and camera motion features". In: Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA, pp. 332–343
20. Kuhmunch C (1997) "On the detection and recognition of television commercials". In: Proc. Int. Conf. on Multimedia Computing and Systems, June 3–6, Ottawa, Canada, pp. 509–516
21. Lehane B, O'Connor NE, Murphy N (2005) "Dialogue sequence detection in movies". In: Proc. Int. Conf. on Image and Video Retrieval 2005, Singapore, pp. 286–296
22. Lienhart R, Pfeiffer S, Fischer S. "Automatic movie abstracting", Universität Mannheim, Reihe Informatik 3/97
23. Lin Y, Lin Y (2005) "Robust face detection with multi-class boosting". In: Proc. Int. Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 680–687
24. Ozkan D, Duygulu P (2006) "Finding people frequently appearing in news". In: Proc. Int. Conf. Image and Video Retrieval, Tempe, AZ, USA, pp. 173–182
25. Petersohn C (2009) "Temporal video structuring for preservation and annotation of video content". In: Proc. IEEE Int. Conf. on Image Processing, Cairo, pp. 93–96
26. Porikli F, Tuzel O, Meer P (2006) "Covariance tracking using model update based on lie algebra". In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, New York, NY, USA, pp. 728–735
27. Satoh S, Kanade T (1997) "Name-it: association of face and name in video". In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, San Juan, Puerto Rico, pp. 368–373
28. Viola P, Jones M (2001) "Rapid object detection using a boosted cascade of simple features". In: Proc. Int. Conference on Computer Vision and Pattern Recognition, Kauai, USA, pp. 511–518
29. Yang J, Yan R, Hauptmann AG (2005) "Multiple instance learning for labeling faces in broadcasting news video". In: Proc. 13th. ACM Int. Conf. Multimedia, Nov, Singapore, pp. 31–40
30. Zhang Yi-Fan, Changsheng Xu, Hanqing Lu, Huang Y-M (2009) Character identification in feature-length films using global face-name matching. IEEE Trans Multimedia 11(7):1276–1288

31. Zhang L, Chu R, Xiang S, Liao S, Li SZ (2007) Face detection based on multi-block LBP representation. Lect Notes Comput Sci 4642:11–18
32. Zhang X, Gaoa Y (2009) Face recognition across pose: a review. ELSEVIER Pattern Recogn 42 (11):2876–2896
33. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: a literature survey. ACM Comput Surv 35(4):399–459

Tobias Schwarze Tobias Schwarze received his diploma of Computer Engineering from the Technical University of Berlin in 2010. During his studies he worked on a couple of projects within real-time vision groups of Siemens Corporate Technology and joined CT Munich for his thesis on semantic video understanding in 2009. Further research interests cover environment perception and human-robot interaction for intelligent autonomous systems.

Thomas Riegel was born in 1960 in Augsburg, Germany. He studied Computer Science at the Technical University of Munich, where he got his Diploma in 1988. Since that time he is with the Siemens Corporate Technology in Munich. He was engaged in a couple of European projects dealing with recovering depth from stereo, surface approximation and interpolation by triangle meshes. Later on he worked on a consistent embedding of synthesized views into virtual worlds for communication purposes.

His current field of activity covers the metadata-based archiving and retrieval of video content. Within that context he manages the according Siemens part in THESEUS. THESEUS is a research program initiated by the German Federal Ministry of Economics and Technology (BMWi) of Germany with the goal of simplifying access to information, combining data into new knowledge and laying the groundwork for developing new services on the Internet.

Seunghan Han received his B.Sc. and M.Sc. degrees in Computer Science from Sogang University, Korea in 2003 and 2006, respectively. Since 2006, he is a Ph.D. candidate at the Institute for Integrated Systems at Technische Universität München, Germany. During this time, he is sponsored by Siemens doctoral scholarship and works in Image Analytics and Informatics at Siemens Corporate Technology in Munich, Germany. His research interests are computer vision and artificial intelligence (knowledge representation and reasoning under uncertainty), system engineering for computer vision and semantic analysis of visual surveillance, multimedia search, retrieval and indexing.

Andreas Hutter Andreas Hutter received his diploma and Dr.-Ing. degrees in communications engineering from the Munich University of Technology in 1993 and 1999, respectively. From 1993 to 1999 he was as a research assistant with the Institute for Integrated Circuits of the Munich University of Technology, where he mainly worked on algorithms for video coding and on the implementation of multimedia systems for mobile terminals. He joined Siemens Corporate Technology in 1999, where he is currently leading the research program for video search and video processing. He has been an active member of MPEG from 1995 to 2006 where he contributed to the MPEG-4, the MPEG-7 and the MPEG-21 standards. He was co-editor of the MPEG-7 Systems standard and he was acting as HoD (Head of Delegation) of the German National Body at MPEG.

He has also been actively involved in several European research projects e.g.in the EU-IST projects ISIS and DANAE as well as in the German BMWi funded project THESEUS.

Stefanie Nowak Stefanie Nowak did her studies in applied computer science with minor media science at the University of Siegen in Germany and finished with a diploma in computer science in 2006.

She joined the Fraunhofer Society in 2005 and started at the Fraunhofer Institute for Digital Media Technology as a PhD student in 2007. She was visiting researcher at the Knowledge Media Institute, Open University, UK in 2009/2010.

Her current research interests focus on evaluation methodologies for visual analysis systems and the automated extraction of semantic information from visual media. She is active in the ImageCLEF benchmarking initiative where she organizes the photo annotation task. In the publicly funded research programme THESEUS, she is concerned with the evaluation of image and video analysis technologies of the core technology cluster.

Sascha Ebel Sascha Ebel was born in 1980 in Berlin, Germany. He studied Computer Science at the Humboldt University of Berlin, with a focus on machine learning and artificial intelligence.

He got his Diploma in January 2010. Since then has been working as a research engineer at the Fraunhofer HHI (Heinrich Hertz Institute) in the Interactive Media-Human Factors group. His main research interests are face detection in images and videos and indexing of multimedia data. At the moment he is engaged in the Theseus project. Within that project he is responsible for the development of a face-detection and tracking system.