Confidence Based Multimodal Fusion for Person Identification

Philipp W. L. Große Hartwig Holzapfel Alex Waibel
InterACT Research, Interactive Systems Labs
Am Fasanengarten 5, Building 50.34
76131 University Karlsruhe, Germany
ph.grosse@web.de {hartwig, waibel}@ira.uka.de

ABSTRACT

Person identification is of great interest for various kinds of applications and interactive systems. In our system we use face recognition and voice recognition from data recorded in an interactive dialogue system. In such a system, sequential images and sequential utterances can be used to improve recognition accuracy over single hypotheses. The presented approach uses confidence-based fusion for sequence hypotheses, for multimodal fusion, and to provide a reliability measure of the classification quality that can be used to decide when to trust and when to ignore classification results.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Experimentation, Reliability

1. INTRODUCTION

In this paper we describe a confidence-based fusion approach for person identification during human-robot interactions. Ekenel, Fischer, Jin and Stiefelhagen [1] have shown that for fusion of audio and video for person identification, adaptive weighting of modalities is one of the primary factors for better recognition results. Their adaptive CRCM approach uses modality weights which are based on the differences of the best two hypothesis scores of each modality. On the other hand Könn, Holzapfel, Ekenel and Waibel [5] had shown that logistic regression can be used to combine different confidence features for good confidence estimation. Our goals therefore were to extend this approach and investigate different confidence features in their suitability. The resulting confidence measures are then integrated in the fusion process, and used as reliability estimates for other system components. The target scenario for this approach is a dialogue system which poses additional requirements, such as online computation of hypotheses instead of observing the complete sequence before a first classification is made.

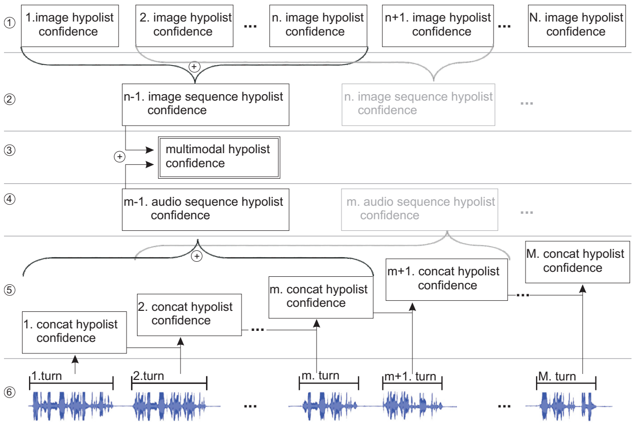

①
1.image hypolist confidence
2. image hypolist confidence
...
n. image hypolist confidence
n+1. image hypolist confidence
...
N. image hypolist confidence
+
②
n-1. image sequence hypolist confidence
n. image sequence hypolist confidence
...
③
+
multimodal hypolist confidence
④
m-1. audio sequence hypolist confidence
m. audio sequence hypolist confidence
...
+
⑤
1. concat hypolist confidence
2. concat hypolist confidence
...
m. concat hypolist confidence
m+1. concat hypolist confidence
...
M. concat hypolist confidence
⑥
1.turn
2.turn
...
m. turn
m+1. turn
...
M. turn

Figure 1: Structure of the multimodal classifier with different classification layers.

In contrast to previous work [5] this approach improves classification accuracy by fusing n-best lists instead of single hypotheses. Using n-best lists allows faster compensation of incorrect hypotheses and facilitates new confidence features from hypothesis lists. As experiments with different features show, good confidence estimation is possible without internal knowledge of the classifier, by just using features that are computed on hypothesis lists. Furthermore, these confidence measures can be used to reliably detect unknown person, as experiments with open set person identification have shown. Open set person identification means that the set of persons includes 'unknown' persons.

2. SYSTEM DESCRIPTION

The system setup conforms to the turn-based dialog system architecture. Each turn corresponds to a dialog utterance by the user during which a single audio file is recorded. Video images are recorded continuously during the dialog. In order to create an assumption about the person in front of the robot several processes have to be passed. Figure 1 gives an overview over the system architecture and its components. The system is divided into six separate subsystems (illustrated through dashed lines): the single image layer (1), the image sequence layer (2), the single turn layer (6), the concat turn layer (5), the audio sequence layer (4) and finally the central multimodal layer (3). Apart from the single turn layer all other subsystems provide hypotheses in the form of n-best hypothesis list and a corresponding confidence. The

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-303-7/08/10 ...$5.00.

additional concat turn layer has been introduced to provide audio utterances adequate for voice ID. Since data collected with the dialog system includes many short utterances (< 1 second), a *concat turn* is simply the concatenation of the single audio utterances, which have been recorded during the dialog.

2.1 Confidence-Based Fusion

Like the title of our paper already indicates we use confidence measures as basis of an adaptive weighting for fusion, which includes fusion of different modalities and per-modality fusion of single hypotheses to obtain sequence hypotheses. In each case, fusion is realized as summation over n-best lists which are weighted by confidence values. This approach is illustrated in figure 1 on several layers, fusion is marked by $\oplus$. Mathematicly spoken the hypothesis of the next higher layer is calculated through:

$$H_{new} = \sum_{i=1}^{N} conf(H_i) \cdot H_i \tag{1}$$

where H respectively denotes an n-best list, and $conf(H_i)$ represents the confidence for this hypothesis. N refers to the *sliding windows* size, i.e. the number of accounted hypothesis lists. In the multimodal case, this function produces a sensible result, even if only one modality is available. Normally, in the multimodal case, N is 2 (audido + video) and each modality contributes a sequence hypothesis n-best list with the respective confidence, which is then merged to a new hypothesis list according to equation 1. For the sequence hypothesis, N denotes the maximum number of accounted single hypotheses. Since the classification approach is designed for a life system, the approach produces a sequence hypothesis starting with the first two single hypotheses and grows the sequence length until N is reached. From there on the sequence is shifted as a sliding window over the single hypotheses with a fixed size.

Each fusion step requires that hypothesis lists are normalized. This is especially important for the multimodal fusion since the hypotheses to be fused originate from different classification methods (k-Nearest-Neighbour and GMMs). We use the following normalization method:

$$\bar{s}_i = \frac{s_i - min}{\sum_{i=1}^{n} (s_i - min)}$$

where s_i denotes the score of the i-th best hypothesis, min denotes the smallest score within the n-best list and n denotes the length of the hypothesis list, in our case n was set to ten. Here, the smallest value is lost, and only influences normalization, because for $s_i = min$ the score gets normalized to zero.

2.2 Confidence Estimation

The term *confidence* in this paper refers to the reliability of the classification. In contrast to the scores of the hypotheses it is calculated on separate features and is a probability value.

Depending on the classification the features which can be used as confidence features are different. For example there are confidence features like the mean gray value (Image) or the approx. distance between the subject and the camera (Dist). Both features are only applicable to face ID classification. Other confidence features like the agreement (Agre) and stability (Stab) can only be used for sequence hypotheses [5]. However, some confidence features could be used throughout the whole system, since they are based on the structure of n-best lists, which are available for both modalities. Those confidence features are the entropy of the n-best list (Ent), the difference between the two highest scores of the n-best list (Diff0), as well as two further difference-based features (Diff1, Diff2).

Like already mentioned the confidence features 'Image' and 'Dist' are image-specific features, for both are - more or less - directly derived from the image data. Both have been shown to be effective for confidence estimation [5].

The two sequence specific confidence features are calculated through correlation of each best hypothesis within the considered sequence. They are suitable for the confidence calculation since they are directly related to the commonness of the agreement (Agre), and respectively to the alternation of the hypotheses (Stab). More precisely 'Agre' denotes the number of single image hypotheses, which are equal to the best hypothesis, divided by the total number of accounted hypotheses (corresponding to the sliding window size). Whereas 'Stab' denotes the number of hypotheses changes relative to the total number of accounted hypotheses within the sequence.

The four n-best list based features (Ent, Diff0, Diff1 and Diff2) are suitable as confidence features because they are directly related to the structure of the n-best list and therefore reflect the probability of confusion. They are calculated as follows:

$$\text{Ent} = -\sum_{i=1}^{N} k_i \cdot log_2(k_i) \qquad \text{Diff1} = \sum_{i=1}^{N} \frac{k_i - k_{i+1}}{i}$$

$$\text{Diff0} = k_1 - k_2 \qquad \text{Diff2} = \sum_{i=1}^{N} \frac{k_i - k_{i+1}}{e^{i-1}}$$

where k_i denotes the score of the i-th best hypothesis, and N denotes the length of the n-best list. It can be seen that the two confidence features 'Diff1' and 'Diff2' are closely related and therefore their values are not statistically independent.

To obtain confidence values in the sense of probability estimates, we use logistic regression [4], to train so-called logit-coefficients. These logit-coefficients are then used to weight the different (confidence) features in the classifier.

3. EXPERIMENTS AND RESULTS

3.1 Data Corpus

Data used for the experiments was collected during dialog experiments in a corridor robot scenario [3], including audio data and video data for multimodal person identification, with the tracking library arthur[1]. For single image face identification and voice identification we use the approaches from Ekenel and Jin, also referenced in [1].

The data has been recorded from a dialog corpus of 38 subjects in 85 sessions. It comprises a collection of single images (recorded at 8 to 15 frames per second) and single audio utterances. The length of recorded sessions varies depending on the dialog length. A session on average contains 1019 single images with 378 face detections and 14 single turns, with a total audio length of 14 seconds. From the perspective of our proposed multimodal system a session on average generated 378 single image hypotheses, 377 image sequence hypotheses, 14 concat turn hypotheses, 13 audio sequence hypotheses and 390 multimodal hypotheses.

[1] http://isl.ira.uka.de/˜nickel/arthur/

Experiments with voice identification have shown that sessions that contain only short audio segments, such as 'yes' and 'no' utterances, didn't contain enough discriminative information to distinguish between speakers with realistic recognition rates. We therefore recorded new audio data for 11 speakers and replaced the original audio data with the new recordings. The new data differs in one important aspect. Though the audio length for voice ID training was just 10 seconds, the recorded utterances were full sentences instead of short utterances.

Given the requirement of independent training and evaluation sets to train the face and voice ID as well as to obtain logit-coefficients for each layer of our system, we echeloned the sets (see figure 2) in such a manner that the used evaluation sets were independent of the corresponding training sets, but never the less could be used another time to train the next higher layer of the system.

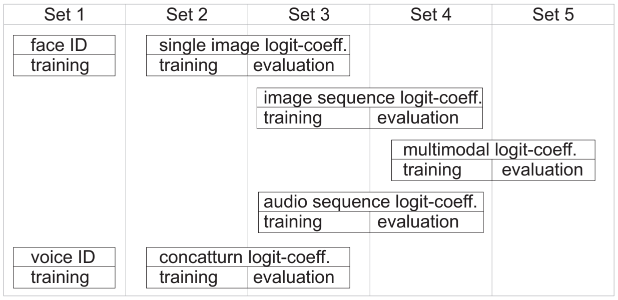


Figure 2: Use of sets for training and evaluation.

So in total five separate sets of data are needed, which are obtained by dispersing the video and audio sessions among the sets. Set 1 was used for face ID and voice ID training, where the face ID was trained on 25 persons and the voice ID was trained on 8 persons. All other sets are used as evaluation data for face ID and voice ID and existed in two versions. The first version, denoted by 'A', contains only those sessions, where all subjects are also contained within set 1 and thus belong to the training set. The second version of the sets, denoted by 'B', contains all sessions from 'A' plus further sessions with 'unknown' persons. For training and evaluating logit-coefficients we used the 'B' versions, since the recognition rate is fairly high and we wanted our system to cope with unknown persons as well. As shown in figure 2, set 2 is used to train logit-coefficients for single image face ID and voice ID. Set 3 is used to train logit-coefficients of sequence hypotheses, set 4 is used to train logit-coefficients of multimodal person ID and set 5 finally is used to evaluate person ID classification on unseen data.

3.2 Selection of Confidence Features

To be able to provide confidences within each layer of our proposed system, we had to select suitable confidence features and calculate logit-coefficients for each of its subsystems. Given the restricted space of this paper we show our approach exemplary on the single image and the image sequence layer.

To obtain suitable confidence features for the single image layer, we calculated single image hypothesis lists for sets 2B and 3B and stored them with all relevant confidence features (Diff0, Diff1, Diff2, Ent, Image and Dist). Afterwards we trained for a wide range of possible confidence feature combinations logit-coefficients based on set 2B and evaluated them on set 3B. Figure 3 shows a detailed section of the corresponding ROC graph. ROC graphs are suitable to compare classifiers which can be evaluated with true positive and false positive rates [2].

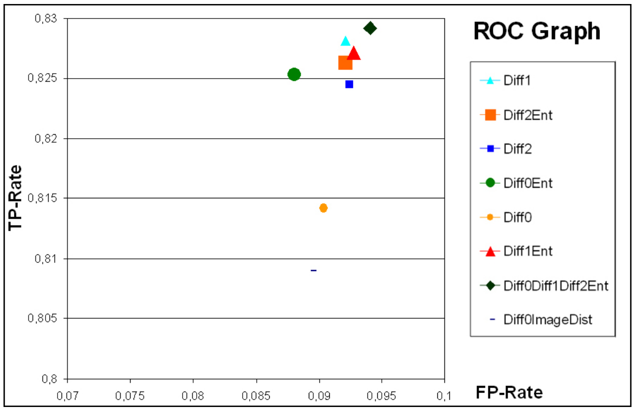


Figure 3: ROC graph showing different confidence features for single image hypotheses (face ID) evaluated on set 3B.

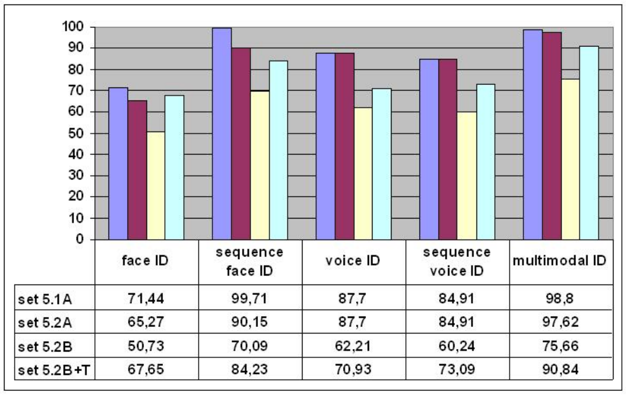
Most confidence features are clustered within the same TP-/FP-Area[2], with a rather low FP-rate (<0.1) and a rather high TP-rate (>0.8), except Entropy alone which has a true positive rate of 0.71. Among the minor differences between the feature combinations, 'Diff0Ent' produces the lowest FP-rate, while being only slightly worse than the best feature combination regarding TP-rate, and thus was used in the final setup. The combination of 'Diff0' and 'Ent' also shows slightly better results than 'Diff0' alone (which was used for the CRCM approach in [1]). The features used in [5] could not fully be transferred to this approach since tracking and face identification methods differ.

To choose suitable confidence features for the image sequence hypotheses we calculated image sequence hypotheses for sets 3B and 4B and stored them together with all relevant confidence features (Agre, Stab, Diff0, Diff1, Diff2, Ent) and again compared different feature combinations within a ROC graph. In this comparison those feature combinations, which took 'Agre' and 'Stab' into account performed best. Before deciding on the best feature combination one has to consider different sequence lengths which is an important aspect of the online system. While it is obvious that with increasing sequence length, the quality of the hypotheses increase, this is not necessarily true for confidence classification. We have calculated possible confidence feature combinations according to the sequence length of 4, 15, 50, 100 and 200, whereof 200 was used in the evaluation. On the given data, 'AgreStabDiff0Ent' shows the highest stability concerning different sequence lengths.

In the following the results of the system and its subsystems are subsumed based on the evaluations of set 5. Figure 4 shows the recognition rates for the different layers and different subsets. Set 5.1A contains only known persons and has been recorded with similar light conditions. Set 5.2A contains only known persons and includes sessions recorded at different points in the building with varying directions of light. Set 5.2B is an extension of set 5.2A, and includes unknown persons, but no unknown classification. Set 5.2B+T is the same data set as 5.2B. Here, a threshold approach has

[2] TP: true positive, FP: false positive

been tested for unknown person detection based on confidences.



	face ID	sequence face ID	voice ID	sequence voice ID	multimodal ID
set 5.1A	71,44	99,71	87,7	84,91	98,8
set 5.2A	65,27	90,15	87,7	84,91	97,62
set 5.2B	50,73	70,09	62,21	60,24	75,66
set 5.2B+T	67,65	84,23	70,93	73,09	90,84

Figure 4: Recognition rates on set 5.1 and 5.2.

The figure shows significant improvements at several layers for known person identification (1st and 2nd bars). Classification of the sequence face ID benefits from competing hypotheses, which in case of false classifications are spread in the Nearest-Neighbor features space. In case of voice ID, the GMM-based classifier tends to produce similar (incorrect) hypotheses. This can also be seen by the best confidence feature which is 'Diff0' and doesn't include 'Agre' and 'Stab'. Thus, sequence voice ID doesn't produce better hypotheses than the concatenated voice ID, but produces better confidence estimation, which is of great importance to the multimodal fusion.

Improvement of the multimodal ID can be seen best on set 5.2A, which has been recorded at more difficult conditions for face identification. Detailed numbers are shown in table 1. On average, the confidences distinguish between correct and incorrect classifications. An exception is seqFaceID, were too few incorrect hypotheses have been seen during confidence training (>99% correct). It can also be seen that unkown persons receive very low confidences, which suggests that unknown classification is possible. The challenge here is to distinguish unknown form incorrect recognition, which can be adressed e.g. by calculating average confidences over a sequence of hypotheses and then applying a threshold. The numbers in figure 4 have been computed with an optimal threshold of 0.3 for the multimodal ID. At this threshold level, 70% unknown was detected correctly, and <0.8% new errors (known vs unknown) are made.

4. CONCLUSIONS

We have presented an approach for confidence based fusion in a multimodal ID classification task. Different features and feature combinations have been investigated regarding their suitability for probability estimation with logistic regression.

All confidence classifiers for the final system make use of distributions of the n-best hypothesis lists. A major benefit of such features is that they can solely be computed on classifier output, without using 'internal' information. The same is true for the features 'Agre' and 'Stab' which cover sequence characteristics. As the experimental results show, the confidence-based fusion approach significantly improves the overall recognition rate.

Together with multimodal hypotheses, on the highest layer, confidences are calculated that can be passed on to other dialog system components. Looking at a sequence of those confidences furthermore allows us to reliably detect unknown persons, even though a single incorrectly classified hypothesis may have a low confidence value as well.

face ID *(Diff0Ent)*			
	number	mean	std. deviation
Set 5.2A hypothesis true	4540	0.487	0.36
Set 5.2A hypothesis false	2416	0.154	0.212
Unknown	1993	0.119	0.166
sequence face ID *(AgreStabDiff0Ent)*			
	number	mean	std. deviation
Set 5.2A hypothesis true	6261	0.491	0.396
Set 5.2A hypothesis false	684	0.432	0.254
Unknown	1988	0.029	0.046
voice ID *(Diff0)*			
	number	mean	std. deviation
Set 5.2A hypothesis true	214	0.452	0.164
Set 5.2A hypothesis false	30	0.336	0.073
Unknown	100	0.315	0.049
sequence voice ID *(Diff0)*			
	number	mean	std. deviation
Set 5.2A hypothesis true	197	0.601	0.284
Set 5.2A hypothesis false	35	0.167	0.145
Unknown	95	0.265	0.196
multimodal ID *(AgreStabDiff0Ent)*			
	number	mean	std. deviation
Set 5.2A hypothesis true	7006	0.804	0.32
Set 5.2A hypothesis false	171	0.326	0.452
Unknown	2083	0.094	0.274

Table 1: Overview of different confidence classifiers.

All confidence approaches generally perform better for face ID than for voice ID on the given data. In the future we think it can be valuable to observe if this is due to different distribution of the generated hypothesis lists or if other 'internal' features of the voice ID should be exploited. Also it would be interesting to explore in which way this approach can be combined with other approaches that explicitly classify unknown persons.

5. ACKNOWLEDGMENTS

This work was realized by OSEO, French State agency for innovation, as part of the Quaero programme. And further more supported by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots".

6. REFERENCES

[1] H. K. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen. Multi-modal person identification in a smart environment. *CVPR Biometrics Workshop, Minneapolis, USA*, June 2007.

[2] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical report, HPL-2003-4, HP Laboratories., 2003.

[3] H. Holzapfel, T. Schaaf, H. K. Ekenel, C. Schaa, and A. Waibel. A robot learns to know people - first contacts of a robot. *KI 2006: Advances in Artificial Intelligence, Springer LNCS*, 4314, 2007.

[4] D. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, 1989.

[5] S. Könn, H. Holzapfel, H. K. Ekenel, and A. Waibel. Integrating face-id into an interactive person-id learning system. *International Conference on Computer Vision Systems (ICVS'07)*, Bielefeld, Germany, 2007.