LECTURER LOCALIZATION AND IDENTIFICATION IN A SMART ENVIRONMENT

Hazim Kemal Ekenel Kai Nickel Rainer Stiefelhagen

Interactive Systems Labs, Universität Karlsruhe (TH)
76131 Karlsruhe, Germany
{ekenel, nickel, stiefel}@ira.uka.de
web: http://isl.ira.uka.de

Abstract

In this paper, an intelligent system that can locate and identify the lecturer in a smart environment is presented. The purpose of the presented system is to provide technologies that can assist the lecturer during his/her presentation and to facilitate lecture retrieval by means of the identity of the lecturer. The system consists of two main building blocks, a robust face detector based 3D position generation -fusing 2D face locations from multiple cameras- and a novel face identification system that utilizes data coming from multiple cameras and video. The performance of the system is tested on the seminar data collected under the CHIL project.

1. Introduction

In our everyday life, electronic devices are becoming to require more time and interest. Mobile phones and computers are two of the best examples of these devices. Efficiently utilizing the time spent on these devices and transforming them from being center of focus to smart systems that can operate unawarely and assist individuals in their daily activities, will be a crucial step in human-computer interactions. With these smart systems, the uni-directional, from-human-to-machine, traditional learning process will become bi-directional in next generation machines. This will be materialized with the invention of machines that can understand, learn the preferences of users and support them in their interactions with the environment or other individuals. CHIL project is aimed to provide technologies and services based on these technologies to achieve this goal. The scenario in CHIL project is a situation in which people interact face to face with people, exchange information, collaborate to jointly solve problems, learn, or socialize, by using whatever means (speech/language, gestures, body posture, data in electronic format, slides, etc.) they choose [1]. The X system monitors the environment and individuals to provide useful services. One of these useful services is to inform the lecturer about the status of the audience, i.e. the attention level of the audience, their contribution frequency to the lecture, etc. The technologies that can provide this service are a targeted audio device –an audio device that can transmit audio only to the intended individual without disturbing the others-, and a system that can locate the presenter's 3D world position. By sending the 3D world coordinates of the lecturer to the pan-tilt unit that the audio device is connected to, the audio signal can be targeted to the lecturer. The other interesting service would be to retrieve the seminar data using the identity of the lecturer. The system presented in this paper consists of a 3D person locator that can find the lecturer and calculates his/her 3D coordinates whenever the service requires to send a message to the lecturer and a face recognizer that can observe the lecturer over a limited video sequence and tries to find his/her identity.

The smart-room used for capturing lectures is equipped with a variety of sensors, including multiple cameras, distant and close-talking microphones as well as microphone arrays. This setup allows us to capture many of the modalities that are typically used by humans (the lecturer as well as the audience) in communication with each other in the given scenario.

For this work, we used four fixed cameras with a native resolution of 640x480 pixels that cover the entire room (see Figure. 1). The lecturer uses a top-mounted video beamer in order to display slides on the whiteboard. The audience is spread all over the room, thus making it difficult to establish a clear spatial separation between speaker and audience area.

The remainder of the paper is organized as follows. In Sections 2 and 3, 2D face detection system and 3D location estimation procedure are explained respectively. In Section 4, the novel face recognition algorithm is presented. Experimental results are discussed in Section 5. Finally, in Section 6, conclusions and future recommendations are given.

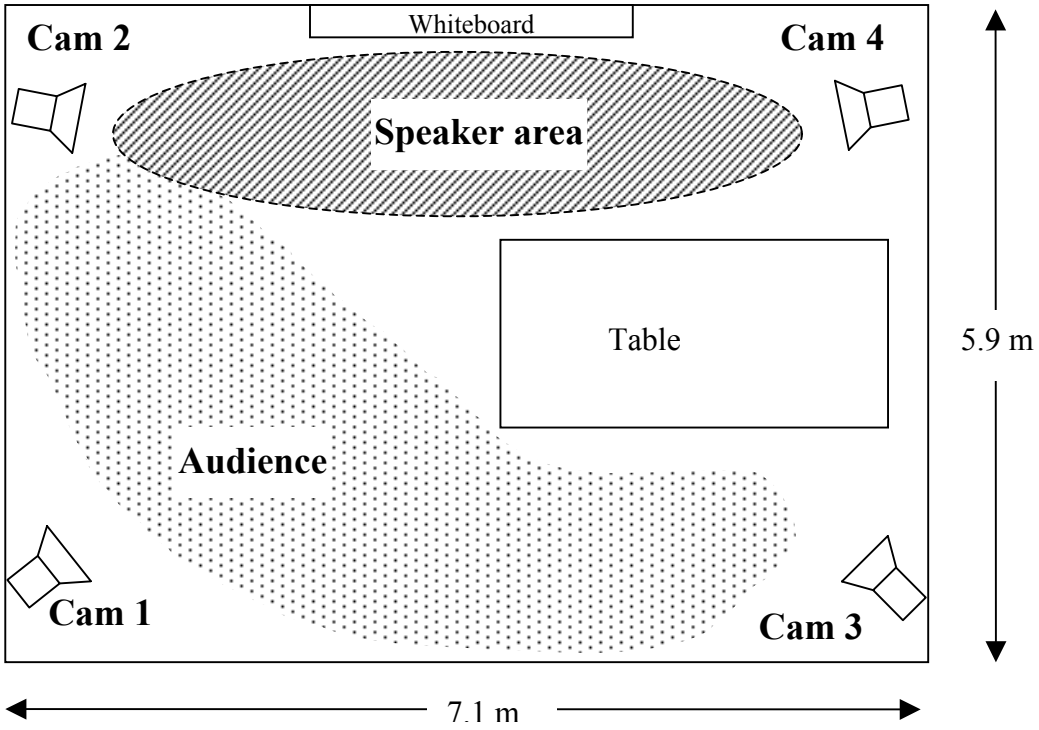

Figure 1. The smart-room is equipped with 4 fixed cameras at a height of approx. 2.7m. The cameras' joint field-of-view covers the entire room. There is no strict separation between speaker area and audience.

Figure 2. Sample camera views from a seminar at the same instant
(Top-left: Camera 1, Top-right: Camera 2, Bottom-left: Camera 3, Bottom-right: Camera 4)

2. Face Detection

The face detection algorithm contains three steps. The first step utilizes prior location distribution to separate the lecturer's presentation area from the audience area. The second step consists of haar-like features [2] based multi-view face detection. The third step is calculation of average skin color likelihoods of each candidate face rectangle provided by the multi-view face detector. The candidate face rectangle with the highest average skin color likelihood is selected as the lecturer's face rectangle.

2.1. Prior Location Distribution

To determine the region of interest (ROI) –the lecturer's presentation area- for limiting the face search region, labelled head center locations in the training set are processed. The original 640x480 resolution image is divided into 64x48 bins and then 2D histogram of the head center locations are calculated. All the bins that have a head center location in it are included to the ROI (in other words, to be included in the ROI, it is sufficient to contain only one head center location). Furthermore, the initial ROI is dilated twice to prevent any misses.

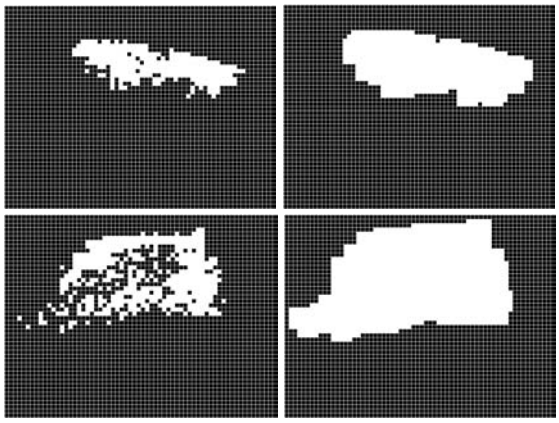

Figure 3. Camera 3 and 4's initial and extended ROIs (top: Camera 3, bottom: Camera 4)

2.2. Multi-view Face Detection

Multi-view face detector is based on the approach presented in [2,3]. For multi-view face detector, two separate -one for frontal and one for profile- face cascades are trained.

Figure 5. Sample training faces –first row: frontal, -second row: profile

2.3. Color Filtering

The skin color samples obtained from the training set is used to build the skin color model. The modelling of the skin color distribution was performed in RGB color space using 3-dimensional histogram.

3. 3D Location Estimation

Given the image coordinates of an object in two or more camera views, it is possible to determine its 3D location. In addition to the image coordinates, this requires knowledge about the intrinsic and extrinsic calibration parameters of the multi-camera setup [4, 5]. After calibration, for each image pixel the line-of-view (LOV) from the camera's projection centre to the depicted object can be calculated. Ideally, the LOV's from the same object in different camera views intersect at the object's true 3D position. In practice, this problem comes down to a least squares solution for the system of equations given by the set of LOV's. The residual error resulting from this calculation can be interpreted as a confidence measure for the validity of the intersection.

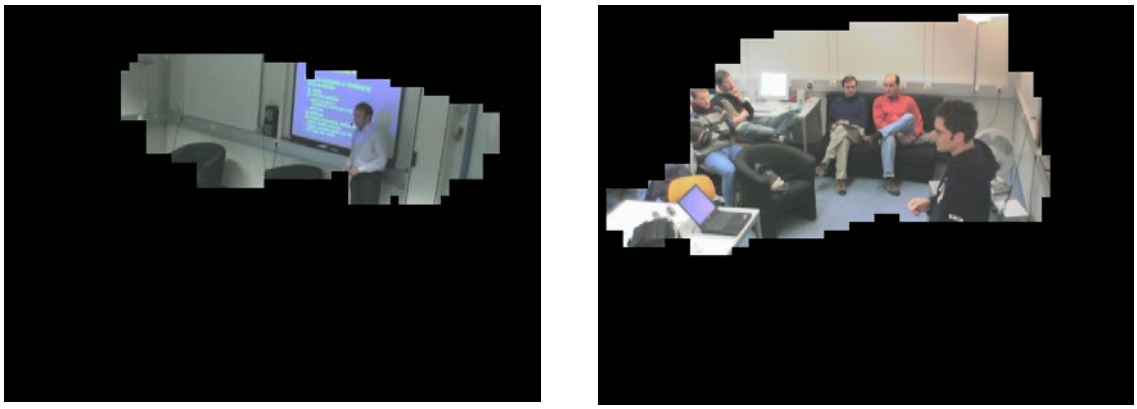

Figure 4. Sample ROI views from camera 3 and 4

Thus, in order to locate the lecturer in 3D, we need to combine the face detector's results (bounding box centroids) from multiple camera views. Due to the positioning of the cameras, and the ability of the face detector to detect faces within a rotation range of +/-90°, the availability of two or more face centroids is generally guaranteed. As Figure 1 shows, the speakers face could simultaneously be found in camera pairs 1&3, 1&2, 3&4 or 2&4. A combined detection in cameras 2&4 is unlikely, as the lecturer generally faces the audience, whereas the combination of directly opposed cameras like 2&3 would not produce a meaningful result.

For each camera view, the face detector outputs one or no face centroid, that either represents the lecturer's face, a face from the audience or no face at all (misses). At this point, we run into a data association problem, because we don't know whether the face in one camera view corresponds to a face from another view. As a consequence, our combination scheme evaluates all possible intersections of the centroids' LOVs in all of the adjacent camera pairs listed above. Considering the fact that intersecting LOVs which do *not* belong to the same physical object results in a high residual error, we select that combination with the lowest residual error, provided that its 3D coordinates are a) inside the speaker area, and b) the height above ground is inside a reasonable range around the average height of a standing person.

4. Face Recognition

Face recognition under unconstrained conditions is still an unsolved problem. A face recognition system should be robust against detection and alignment errors, less sensitive to illumination and background variations and easily extendible to detect and to recognize unknown people. Furthermore, it should naturally weight the contributions of the frames from multiple cameras and video sequence to face classification. With these requirements in mind, a novel face recognition algorithm is developed. To provide less sensitiveness to illumination and background variations the face appearance is modelled locally. That is, the detected and resized face is divided into 8x8 pixels resolution blocks and each block is represented with discrete cosine transform (DCT) coefficients. Although, the paradigm of local appearance-based face recognition can also benefit from other data dependent or independent basis functions, data independent basis are preferred to use, since there is no alignment step involved during training for extracting proper basis as in the case of principal component analysis (PCA). DCT is chosen for its compact representation capability, fast computation and data independent nature. To provide robustness against detection and alignment errors, artificial samples are generated from the original training face images by translating and scaling them (the artificial data generation process is not limited to only translation or scale, it can provide any other variations like illumination, face rotation, view morphing, etc.). To increase discrimination between candidate individuals and to provide robustness against false detections (detecting background as a face), in the classification step, class-specific linear discriminant analysis (CS-LDA) is performed. That is, each class has its own Nx1 projection vector, where N denotes the size of the feature vector. By performing CS-LDA, the multi-class classification problem becomes a two-class classification. The training data for genuine class consists of samples from the true candidate, whereas the training data for impostor class consists of the other people's samples plus random background samples. The distribution of projected genuine and impostor data in one-dimensional space is modelled with univariate Gaussians. The decision is taken by applying Bayes rule.

$$P(C_{k,1} \mid x) = \frac{P(x \mid C_{k,1})P(C_{k,1})}{\sum_{i=1}^{2} P(x \mid C_{k,i})P(C_{k,i})}$$

where $C_{k,1}$ denotes genuine class and $C_{k,2}$ denotes impostor class of the k^{th} individual. $P(C_{k,1})$ and $P(C_{k,2})$ are taken 0.5. If $P(C_{k,1} \mid x)$ is bigger than 0.5, than the test image is assigned to k^{th} individual. From the equation above, there may be three cases observed. In the first case, for every "k", $P(C_{k,1} \mid x)$ may be smaller than 0.5. This can occur either from a background sample detected as a face or from an unknown face (At the moment, since close-set classification is performed, the cases of false detection and unknown people are combined, otherwise it can be easily separated by performing a hierarchical scheme. For example, at first, the rectangle can be examined to see whether it's a face or not and afterwards, it can be classified as known or unknown person). In the second case, there may be more than one "k" such that $P(C_{k,1} \mid x)$ is bigger than 0.5. In this case the most probable candidate can be selected. In the third case, which is the ideal case, only one of the individuals has $P(C_{k,1} \mid x)$ bigger than 0.5. The extension of this system to multiple camera and video schemes is quite easy and natural –accumulate the $P(C_{k,1} \mid x)$s that are bigger than 0.5 over multiple cameras and video sequence. By this, the video data and multiple camera information is utilized naturally, and good frames –the frames which contains properly detected, high resolution images- are inherently weighted more. The overview of the algorithm is as below:

Training:
1. Artificially generate data to account for improper detection and alignment of faces or illumination variations, etc.
2. Perform local appearance modelling. Perform DCT on 8x8 pixels blocks and choose only the DCT coefficients that contain more information by zig-zag scanning the DCT image.
3. Perform class-specific projection for each class.
4. Model the distribution of projected genuine and impostor data with univariate Gaussians.

Testing:
1. Perform local appearance modelling (as in training step 2).
2. Project feature vector to each class.
3. Use Bayesian rule to obtain the matching score for each class.
4. Accumulate scores over multiple cameras and video whenever the score is bigger than 0.5.
5. Choose the candidate with the highest score.

5. Experiments

The presented system for the localization and identification of a lecturer has been evaluated as part of the annual perceptual technologies evaluation campaign of the X project. In this evaluation, several vision- as well as audio/speech-related tasks are addressed by a number of participants of the X project and are also open to external participants. The visual evaluation tasks include: face and head detection, 3D person tracking, face recognition, head pose estimation, hand tracking and pointing gesture recognition.

5.1. Data Set

The evaluation data consist of 7 seminars recorded in the XYZ Lab. of the University of X. in 2003. Each seminar is given by a different lecturer. These seminars are recorded in five different dates. There are one week to one month time gaps between different recording dates. The lectures in the same day are recorded consecutively. Each seminar is divided into 4 sparse, non-overlapping segments of 5 minutes length. From these four segments, the last two are used for training and the first two are used for testing.

To obtain ground truth for the evaluation, the centroids of the lecturer's head on every 10^{th} frame in each of the video sequences from the different cameras were labelled. From these head centroids the lecturer's 3D position in the room was computed using the calibration information available for all cameras. This 3D position was used as ground truth for the 3D person localization task.

5.2. Person Localization Results

In the face detection task, we used manually labelled around 2400 frontal images and 2700 profile images for training the multi-view face detector. Similarly, for skin color distribution modelling, skin segments are cropped from the labelled faces. At the end, a global skin color distribution model and features for multi-view face detection are obtained. No seminar-specific or camera-specific information is used.

The results of face detection can be seen from Table 1. The measure for correct detection is the Euclidean distance between found head centroid and labelled head centroid. The threshold value is determined to be 15 pixels, which is approximately the half of head size, averaged over all cameras. The face detector's performance is tested on 16703 frames. As can be observed from the table, the correct detection rates for camera 2 and 4 are lower than camera 1 and 3. The main reason for this performance difference is that often other people than the lecturer can be seen in the region of interest (see Figure 4, right picture). Therefore, there may be instants where a background person is selected as a lecturer. Another reason might be the higher resolution of the faces captured by camera 2 and camera 4 (they are closer to the lecturer) such that the threshold is relatively low for them. Misses are caused generally because of too bright or too low illumination and views containing no features –views from behind of the lecturer where only his/her cheek is visible.

	Success (%)	**False detections(%)**	**Misses (%)**
Camera 1	91.88	3.8	4.32
Camera 2	79.78	14.04	6.18
Camera 3	91.47	1.49	7.01
Camera 4	73.99	14.76	11.25

Table 1. Face detection results

The results of the 3D lecturer localization as described in Section 3 can be seen in Table 2. In order to evaluate the system, the manually labeled head centroids from all 4 camera views were transformed into one 3D-coordinate which then can be compared to the 3D-hypothesis. The error is defined as the average Euclidean distance between these two points.

Because a correct face detection in 2 camera views is required for the generation of a 3D-hypothesis, it is obvious, that a face not being detected leads to a missing 3D hypothesis. This is why the percentage of misses in Table 2 is close to the average of misses in Table 1.

Subject	Avg. err (m)	Misses (%)
1	0.10	1.6
2	0.07	6.8
3	0.43	15.2
4	0.11	4.7
5	0.12	3.4
6	0.28	9.8
7	0.17	3.5
Average	0.18	6.3

Table 2. Results of 3D face localization

5.3. Face Recognition Results

In the face recognition task, for each lecturer, five frames that contain their frontal face images are selected from segments 3 and 4 for training. For testing, 20 uniformly sampled, non-overlapping 100 frames sequences from each camera are selected from segments 1 and 2. The classification is performed over these 100 frames. The detection of faces is done automatically. The reason for preparing this face recognition scenario is to classify the seminars with respect to possible lecturers, whose frontal polaroid pictures are available in the database.

Face recognition is performed on all rectangles provided by the multi-view face detector. No average skin color likelihood of the provided rectangles are examined. The reason is that the face recognition system is expected to be robust against false detections. In face recognition experiment, 66.2% average correct classification rate is obtained. Although this result may seem low at the first sight, when the task of single-view-to- multi-view face recognition, the high variability of illumination conditions on the lecturer's face both due to the projector's beam and illumination sources in the room, low resolution (<30 pixels –30 pixels is the head's resolution-), are taken into consideration, it is quite promising. If the correct recognition rate for each individual is examined separately, then it can be observed that 100% can be achieved for some individuals. Also some very low recognition rates can be seen in the table: 0%, 30%. These results may occur due to insufficient proper frames in the video sequence. Other reasons for the obtained results can be the identity candidates (the lecturer listening the other lecturer's seminar) in the audience, poor background model for building the impostors, and missing artificially generated illumination variations.

6. Conclusions

In this paper, a 3D person locator and a person recognizer is presented. Both of the systems use face data as input to derive information. The approaches used in the systems are explained briefly and the experimental results are given for each of the developed technologies. Very satisfying results are obtained from the 3D person locator: an average error of 18cm for the lecturer's head centroid suffices for a variety of possible applications. Face recognition results are also promising regarding to the difficulty of the conditions and algorithm's generic, flexible structure. As a future study, the replacement of 3D person locator system with a continuous 3D person tracking system is planned. In addition, the developed generic face recognition system will be tailored considering the requirements of this specific scenario, it will be also integrated to the tracking system to eliminate background people that mislead the system.

7. References

[1] CHIL project website, http://chil.server.de/.

[2] R. Lienhart, J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", IEEE ICIP, 2002.

[3] M. Jones, P. Viola, "Fast Multi-view Face Detection", IEEE Conference on Computer Vison and Pattern Recognition, June, 2003.

[4] J.Y. Bouguet, Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/

[5] Z. Zhang, Flexible Camera Calibration by Viewing a Plane from Unknown Orientations, *International Conference on Computer Vision (ICCV'99)*, Corfu, Greece, pages 666-673, September 1999.

Figure 6. Sample face detection outputs