THE NOLDUS DATABASE: AUTOMATED RECOGNITION OF RESTAURANT RELATED ACTIVITIES FOR THE RESTAURANT OF THE FUTURE

Moritz Kaiser, Dejan Arsić, Benedikt Hörnler, Martin Hofmann, Gerhard Rigoll

Institute for Human Machine Communication
Technische Universität München
Arcisstr. 21, 80333 München, Germany
moritz.kaiser@tum.de

ABSTRACT

The 'Restaurant of the Future' is a futuristic approach for consumer research in food-related scenarios. Customers are closely monitored by cameras. Currently the video footage is evaluated and annotated manually. For a gradual automation of this process a video database with eating scenarios is published. The database is quite challenging with light changes, overexposure, occluded objects, and equally colored objects, such as glass, plate, skin, and cloth. Basic activities, as drinking, eating, coming in, going out, etc. are manually annotated and can be used as ground truth. A baseline system, consisting of five recognition modules and a rule-based activity recognition method is presented in this treatise. The results can be used as baseline results for comparison within the community.

Index Terms— Multimedia databases, object recognition, behavior recognition, tracking, restaurant of the future

1. INTRODUCTION

Currently almost 90% of the new food products that are launched disappear from the market within one year. One of the main reasons for this high failure rate is the underestimation of the role that situational factors play in food choice. Thus, companies are pushing to find methods to analyze thoroughly the food-related behavior of consumers. In an attempt to satisfy the food suppliers' need for proper consumer analysis Noldus Information Technologies started a pilot project called 'Restaurant of the Future'. Various notable food companies are partners of this project. Among many measuring instruments and sensors, numerous video cameras are installed to monitor the customers. However, currently video recordings have to be annotated and evaluated manually. In order to gradually automate the activity recognition it is planned to analyze their behavior in a similar way as video surveillance systems operate at the moment [1, 2, 3, 4]. Although acoustic cues might also be helpful to characterize activities [5], these have been dismissed due to privacy reasons.

Thus, we identified a strong need for a novel database for eating scenarios on whose basis it is possible to evaluate and compare new approaches and technologies. We recorded the Noldus Database with eleven videos, each lasting approximately 4 minutes, which will be available for the community on request. Several characteristics make the database quite challenging, such as changing light, overexposure, equally colored objects, skin, and cloth, and objects that are occluded by the hands. For the Noldus Database we observed very low performance of state-of-the-art methods, such as the Viola and Jones face detector [6] and current skin color detectors [7], which is a further indicator that the database is demanding. Certain important activities such as eating, drinking, coming in, sitting, going out are manually labeled and delivered with the videos as ground truth for proper evaluation.

An activity recognition system well-suited for the Noldus Database is presented. It is based on five simple recognition modules that show stable performance for these particular eating scenarios. The low level information from the recognition modules is evaluated by a rule-based activity recognition method and remarkable results are achieved. These results can be used as baseline results for further, more sophisticated approaches.

The paper is structured as follows. In Sec. 2 the Noldus Database is explained in detail and particular difficulties of the database are discussed. The five recognition modules of our activity recognition system are presented in Sec. 3 and in Sec. 4 a method for rule-based activity recognition is outlined. Baseline results are given in Sec. 5 followed by a short conclusion in Sec. 6.

2. THE NOLDUS DATABASE

Due to the lack of public data containing restaurant related behavior, Noldus Information Technology recorded their own database. It is available for download on request at the authors. A total of 11 persons have been instructed to simulate a visit to a restaurant. The recordings took place in one of the laboratories of the 'Restaurant of the Future'. The test subjects would sit down at a table and look straight into a PAL resolution camera, mounted in front of the table, as illustrated in Fig. 1. The table has been either already set up, containing a plate of soup, a spoon, and a glass of milk or juice, or

Fig. 1. Screenshots from the Noldus Database. The camera is mounted in front of the table and the persons have a drink and eat soup.

the dishes were brought to the table by a waiter. In any case, first of all an empty scene is recorded, allowing us to build a background model. After entering the scene, the person could behave as he or she wanted, and chat, wait and relax, eat the soup, take a sip from the glass or drink the entire content. All eleven participants were instructed to behave as normal as possible, to allow realistic movements.

While the scenario setup seems rather simple at the first glance, various difficulties have been included. Due to the inconvenient illumination, parts of the faces are overexposed. Most face detection algorithms cannot detect just half of the face, as the other half is almost white. Furthermore, it is almost impossible to use standard skin color detection algorithms, due to the large variance of the present skin color. Therefore, other algorithms have to be developed in future to reliably detect hands and faces within the Noldus Database. A further challenge is the detection of dishes. As these have varying form and color it is hard to create a generative model [8]. Further difficulties in tracking appear as the dishes are frequently covered by the person's hands or change their color as the glasses' content is vanishing. Color based methods are also handicapped, as in some cases all of the dishes are white or can be even confused with overexposed skin regions or the clothing of the person.

The entire database currently contains eleven videos with different persons, is gender balanced, and age varies from 25 to 54 years. Each of the videos is approximately four minutes long, resulting in a total of roughly $68,000$ frames. For proper evaluation of existing and novel algorithms the entire database has been labeled. A new labeling tool has been implemented, allowing labels on frame level. It has been assumed that a person can perform one of the following activities: Coming in, sitting, drinking, eating, going out, or no person. Other objects, namely the plate and the glass, that could also be interesting for potential applications are annotated with one of the following states: No object, moving, or on table. The ground truth is stored in the XML format and will be delivered with the videos.

3. RECOGNITION MODULES

The task of activity recognition is divided into several subtasks which are performed by five simple recognition modules.

3.1. Face Detection and Tracking Module

For face detection and tracking the Viola and Jones face detector [6] (currently one of the best face detectors [9]) has been tested. However, results were unsatisfying. Thus, a less complex approach was chosen. Since it is assumed that the camera is fixed, as a first step a background subtraction is performed. In order to prevent shadows from being detected as foreground, we applied the following formula:

$$FG(x,y) = \begin{cases} 1 & \text{if} \quad \theta_{\text{low}} \leq \frac{I_V(x,y)}{B_V(x,y)} \leq \theta_{\text{up}} \\ & \quad \cap \ |I_S(x,y) - B_S(x,y) \leq \theta_S| \\ & \quad \cap \ |I_H(x,y) - B_H(x,y) \leq \theta_H| \\ 0 & \text{else} \end{cases} \tag{1}$$

where we chose $\theta_{\text{low}} = 0.8$ and $\theta_{\text{up}} = 1.2$ to account for shadows and brightening. The indices H, S, and V stand for the color components in the HSV color space, B is the background image, I the current image, and FG the binary foreground mask computed for the current image.

As a second step skin color is detected. State-of-the-art skin color detectors ([10], [11], and [12]) have been tried with the Noldus Database and failed. The main reasons are overexposure, huge variance of skin color, and objects and wall having colors similar to the skin color. Therefore, we built a skin histogram for this particular setting with eight small skin templates. If the system is used for another setting the eight skin templates can be easily replaced. The templates were transformed into the HSV color space and the entries of the histogram is computed:

$$M(h,s,v) = \sum_{x,y} \begin{cases} 1 & \text{if } T_H(x,y) = h \cap T_S(x,y) = s \\ & \quad \cap \ T_V(x,y) = v \\ 0 & \text{else} \end{cases} \tag{2}$$

where T is the skin template. Each of the three components of the HSV color model has 20 bins in the histogram. For each pixel $I(x,y)$ in the current image the skin color probability can be approximated by

$$P\big((x,y) \text{ is skin}\big) = \frac{M(I_H(x,y), I_S(x,y), I_V(x,y))}{N_{\text{pix}}}, \tag{3}$$

with N_{pix} being the number of template pixels that have been used to create the histogram. The face is considered as detected when there is a sufficiently large amount of skin in the upper half of the image. Subsequently, it is tracked with the robust camshift tracker [13]. The camshift tracker takes the skin probability image as input. The location, the size and the orientation of the face is computed according to the formulas in [13].

3.2. Hand Detection and Tracking Module

For the hand detection and tracking module the same approach as for the face detection and tracking is applied. However, hand tracking is quite unstable as many small snippets of skin color are visible, e.g. at clothing or objects on the table. Thus, we multiply the skin color probability image with a binary mask. For the right hand the binary mask is 1 for each pixel except for the following regions: The left side of the image determined by the center of the face bounding box, background, and any object (glass or plate). The mask for the left hand is defined analogously.

3.3. Table Detection Module

The table detection module is run only once at the beginning. All areas that have the same color as a sample of the table are detected via the flood fill algorithm [14], as shown in Fig. 2(b). Subsequently, a rectangle with edges parallel to the image borders is found with a least-squares minimization, as depicted in Fig. 2(c). The rectangle is matched so that it covers as much of the table pixels as possible while covering as little of non table pixels as possible. With the table rectangle the detection of the hands, the glass, and the plate can be significantly speeded up, since the region of interest (ROI) can be set to this rectangle.

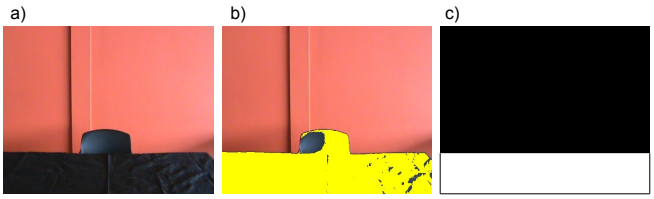

Fig. 2. Table detection module: (a) Background image. (b) Table area detected via flood fill. (c) A rectangle is fit to the table area.

3.4. Glass Detection and Tracking Module

In order to detect the glass on the table the module reads in a template of the glass. The glass is searched only on foreground within the table rectangle. For each location a part of the table region with a size equal to the template's size is cropped out and compared to the template. As a measure of equality the correlation coefficient is taken which is computed by

$$CC_{\text{norm}}(x,y) = \frac{\sum_{x',y'} \left[T(x',y') \cdot I(x+x',y+y')\right]^2}{Z(x,y)}, \quad (4)$$

where T is the template and I the current image. The normalization coefficient is

$$Z(x,y) = \sqrt{\sum_{x',y'} T^2(x',y') \cdot \sum_{x',y'} I^2(x+x',y+y')}. \quad (5)$$

If the greatest correlation coefficient is above a certain threshold (in our case 0.8), the glass is detected and the tracking starts.

For tracking the camshift tracker which was already employed for face tracking is applied. A histogram with 30 bins is learned from the hue component of the template. With this histogram the probability image that the camshift tracker requires as input can be generated. Background pixels are set to zero. Note that only pixels that satisfy the constraints

$$55 \leq T_S(x,y); \quad 65 \leq T_V(x,y) \quad (6)$$

are considered for building up the histogram, since otherwise the hue component is too unreliable. If there are too few pixels satisfying the constraints, the adaptive template tracker that is described in the next section is used.

3.5. Plate Detection and Tracking Module

The plate is usually neither moved too much nor rotated. Thus, a tracker based on an adaptive template [15] was chosen that is less flexible in terms of rotation but more stable if the object stays in the same pose. A plate template is saved and a match is searched on foreground within the table rectangle. The normalized correlation coefficient (Eq. 4) is computed and if $CC_{\text{norm}} \geq 0.8$ the plate is found. Then for each frame the location with the highest CC_{norm} in a 30×30 neighborhood around the current location is searched. When a certain time (in our case 60 seconds) of eating activity has been observed, the template is updated with the current plate.

4. ACTIVITY RECOGNITION

In order to perform activity recognition the information obtained by the five simple recognition modules are gathered and evaluated. We determined three objects of interest: the person, the glass, and the plate. All of those objects of interest can be modeled as finite state machine. The person can be in one of the six states depicted in Fig. 3(a). A glass can be in the one of the three states illustrated in Fig. 3(b). The numbers in Fig. 3 mark the state transitions. Table 1 and 2 list the transition conditions that trigger the state transitions for the person and the glass, respectively. The observation of the plate is implemented analogously to the glass. Figure 4 shows a screenshot from our activity recognition system. The current states of the three objects of interest are depicted in the upper left corner.

Sometimes, misinterpretation of the scenario happens. We established the following safety rules, in order to prevent a corruption of the results for a whole sequence only because of one mistake:

1. Glasses must not be too long off the table
2. A person cannot drink too long
3. If the person drinks, the hand tracker is switched off

The main output of the system is a text file with actions that can be easily compared to the ground truth values available with the videos.

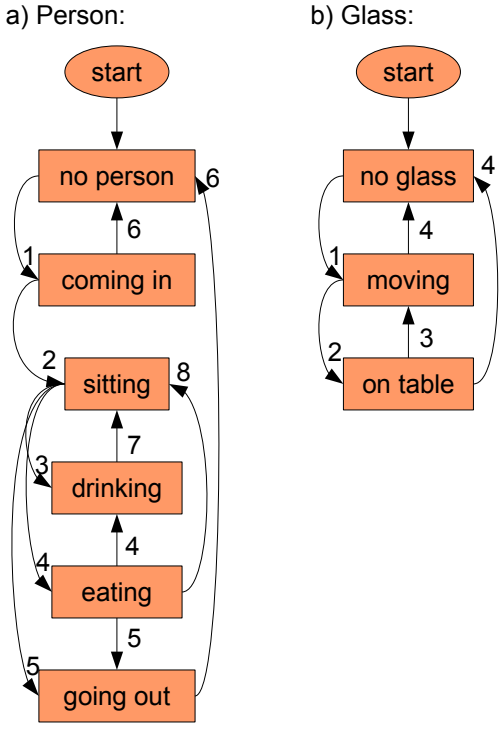

Fig. 3. State diagram of (a) person and (b) glass.

Transition	Transition condition
1	More than 10% of the image is foreground
2	Face is detected and still for 3 seconds
3	Glass and face overlap
4	Hand and face overlap
5	Face moves quickly in y-direction
6	Less than 10% of the image is foreground
7	Glass and face do not overlap
8	Hand and face do not overlap

Table 1. Transition conditions for the person.

Transition	Transition condition
1	Glass detected
2	Glass still for 3 seconds
3	Glass not still for 3 seconds
4	Camshift: Glass gets too big or too small Adaptive template tracker: CC_{norm} is too small

Table 2. Transition conditions for the glass.

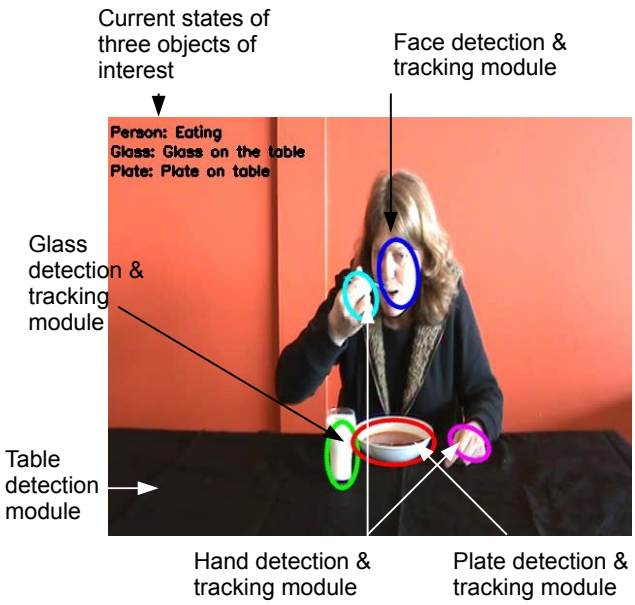

Fig. 4. Screenshot of the proposed activity recognition system.

[%]	Glass	Plate	Person
Correct Frames	83.00	91.80	83.16
Wrong Frames	17.99	8.20	16.84
Best sample	98.41	97.82	91.20
Worst sample	59.34	63.97	74.63

Table 3. Detection results for glasses, plates and persons.

5. BASELINE RESULTS

The simple activity recognition performed by our system has been tested with the Noldus Database. The evaluation can be used as baseline results for more sophisticated recognition systems. Table 3 shows the percentage of frames that could be detected correctly, which we will refer to as *correct recognition rate*. Although the data is challenging, the results are remarkable. It can be seen that the plate detection and tracking module works extremely robust and achieves a correct recognition rate of 91.80%. Even though the task of tracking the glass is more demanding, because it is rotated when the person drinks, it is occluded by the hand, and changes its appearance when it gets empty, still a correct recognition rate of 83.00% can be achieved. For the person, that can be in much more states, the correct recognition rate is 83.16%. This is a surprisingly high detection rate considering the simplicity of the applied classifiers. Nevertheless the variance in-between videos has to be considered, as the algorithms reach by far less in a few obviously harder cases. Commercial applications can therefore not yet be performed and further research is necessary.

6. CONCLUSION

Consumer research for nutrition, as conducted in the 'Restaurant of the Future', currently relies on manual labeling of video data. In order to automate this process the Noldus Database is presented that can be used as a benchmark database for activity recognition in eating scenarios. The database consists of 11 videos, each lasting approximately 4 minutes. Activities are annotated for each frame and this ground truth is also available. A simple activity recognition system is presented, consisting of five recognition modules and a rule-based activity recognition method and it shows good performance on the database. The results of our system can be used as baseline results for further algorithms.

7. REFERENCES

[1] Dejan Arsic, Björn Schuller, and Gerhard Rigoll, "Suspicious behavior detection in public transport by fusion of low-level video descriptors," in *ICME*, 2007, pp. 2018–2021.

[2] Alexander Artikis and Georgios Paliouras, "Behaviour recognition using the event calculus," in *AIAI*, 2009, pp. 469–478.

[3] Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and José M. Molina, "Creating human activity recognition systems using pareto-based multiobjective optimization," in *AVSS*, 2009, pp. 37–42.

[4] Liang Wang, Tao Gu, XianPing Tao, and Jian Lu, "Sensor-based human activity recognition in a multi-user scenario," in *AmI*, 2009, pp. 78–87.

[5] B. Schuller, M. Wimmer, D. Arsić, T. Moosmayr, and G. Rigoll, "Detection of security related affect and behaviour in passenger transport," in *Interspeech*. July 2008, pp. 265–268, ISCA.

[6] Paul A. Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proceedings Graphicon 2003*, 2003.

[8] Constantine Papageorgiou and Tomaso Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[9] Darryl Greig, "Video object detection speedup using staggered sampling," in *IEEE Workshop on Applications of Computer Vision (WACV)*, December 2009, pp. 23–29.

[10] Chris Stauffer and W. Eric L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999, pp. 2246–2252.

[11] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696–706, 2002.

[12] Peter Peer, Jure Kovac, and Franc Solina, "Human skin colour clustering for face detection," in *EUROCON*, September 2003, vol. 2, pp. 144–148.

[13] Gary R. Bradski, "Computer vision face tracking for use in a perceptual user interface," Tech. Rep. Q2, Intel Technology Journal, 1998.

[14] John R. Shaw, "An efficient flood fill algorithm," May 2005, Online article. http://www.codeproject.com/gdi/QuickFill.asp.

[15] Protik Maitra, Stan Schneider, and Min C. Shin, "Robust bee tracking with adaptive appearance template and geometry-constrained resampling," in *IEEE Workshop on Applications of Computer Vision (WACV)*, December 2009, pp. 443–448.