Detection and Tracking of Facial Features in Real Time Using a Synergistic Approach of Spatio-Temporal Models and Generalized Hough-Transform Techniques

A. Schubert
Institut für Systemdynamik und Flugmechanik,
Fakultät für Luft- und Raumfahrttechnik,
University of German Armed Forces Munich
D-855777 Neubiberg
Germany
e-mail: Andreas.Schubert@unibw-muenchen.de

Abstract

The proposed algorithm requires the description of the facial features as 3D-polygons (optionally extended by additional intensity information) which are assembled in a 3D-model of the head provided for in separate data files. Detection is achieved by using a special implementation of the Generalized Hough-Transform (GHT) for which the forms are generated by projecting the 3D-model into the image plane. In the initialization phase a comparatively wide range of relative positions and attitudes between head and camera has to be tested for. Aiming for illumination independence, only information about the sign of the difference between the expected intensities on both sides of the edge of the polygons may be additionally used in the GHT. Once a feature is found, further search for the remaining features can be restricted by the use of the 3D-model. The detection of a minimum number of features starts the tracking phase which is performed by using an Extended Kalman-Filter (EKF) and assuming a first or second order dynamical model for the state variables describing the position and the attitude of the head. Synergistic advantages between GHT and EKF can be realized since the EKF and the projection into the image plane yield a rather good prediction of the forms to be detected by the GHT. This reduces considerably the search space in the image and in the parameter space. On the other hand the GHT offers a solution to the matching problem between image and object features. During the tracking phase the GHT can be further enhanced by monitoring the actual intensities along the edges of the polygons, their assignment to the corresponding 3D-object features, and their use for feature selection during the accumulation process. The algorithm runs on a Dual Pentium II 333 MHz with a cycle time of 40ms in real time.

1 Introduction

The method presented is part of a system for determining and tracking human gaze direction and for monitoring attention which is part of a cockpit assistance system. It is designed mainly to detect pilot errors and support him in the task he is paying attention to [14], [12]. The subsystem for gaze detection uses among other hardware equipment a pan and tilt camera platform arranged approximately 80 cm in front of the person to be monitored. Mounted on the platform are a wide angle camera mapping the pilot's head on the central third of the image (horizontally) and a tele-camera which focuses the eye region. For the task of gaze detection from video sequences only, accurate determination of head position and attitude is indispensable, e.g. [10].

Apart from the application mentioned above, the determination of gaze direction as well as face detection and tracking is important in a large number of other applications, most notably in human machine interaction, e.g. [7].

2 Related Literature

The task of face detection and tracking has been treated by a large number of publications and it is impossible here to give a complete overview (see e.g. [19]). Sometimes, color based approaches are used, e.g. [17] while in other publications 2D approaches are employed (e.g. [2]). 3D-approaches for applications concerning faces are becoming more prevalent (cf. e.g.[19]) and in some publications 3D-models are used for face tracking, e.g. [9]. However, none of the approaches uses explicit 3D models for face and facial features during tracking and detection nor are facial features represented by polygons consisting of 3D line

which can be extracted by edge detection. The use of line features offers several advantages. They are illumination independent and they can be accurately projected into the image plane [5] taking account physical constraints. Other approaches for feature detection and tracking like correlation based techniques, e.g. [9], or wavelet transforms do not take into account transformations due to all spatial degrees of freedom. Furthermore, by use of generic models, a general description can be adapted conveniently to a special case by appropriate choice of parameters, e.g. [5].

The approach proposed relies in essence on four basic elements. One is the well known Generalized Hough Transform (GHT) while another is the use of spatio-temporal models which can be applied efficiently in Extended Kalman-Filters (EKFs). Two more elements are a hierarchical scene representation and a fast and efficient method for edge detection.

Originally developed for detecting lines in images the Hough Transform has first been extended to other primitive features and finally for detecting arbitrary 2D shapes (GHT) [1]. In [16] it has been employed to detect simple 3D objects in 2D images under various aspect conditions by detecting basic features (e.g. lines), calculating possible aspect ratios and incrementing the corresponding accumulator arrays. Among its main advantages are the GHT's relative robustness against noise, partial occlusion, and multi-object environments all of which have to be accounted for in real images. It's main drawbacks are considerable computational and storage requirements which has lead to a variety of methods trying to reduce the associated parameter space, e.g. [11]. The approach presented here uses spatio-temporal models to reduce the parameter space as well as the image space from which the edge elements are extracted.

Spatio-temporal models have been successfully employed in various computer vision applications ranging from vehicle guidance, e.g. [4] to grasping of a free floating object in orbit, e.g. [6], and recently to human gaze detection, e.g. [14]: On one hand they consist of 3D geometrical models borrowed from the domain of computer vision, e.g. [3]. Considering only line features, each relevant object is assigned a coordinate system relative to which object points and (by linking the latter) lines are defined. Subobjects may be assembled to objects and objects are represented in a scene using a hierarchical representation called scene tree which also stems from computer graphics. The links between objects of the scene tree are assigned homogeneous transformation matrices (HTM) describing the relative position and attitude between objects. Cameras are also incorporated in the scene representation for which a pinhole camera model is provided thus enabling projection of the objects into the image plane. On the other hand spatio-temporal models include dynamical models. This is achieved by assigning degrees of freedom (state variables) to the objects which concern primarily the relative position and attitude between objects but also parameters describing their shape. Principally, linear discrete dynamical models are used.

Spatio-temporal models may be employed efficiently for the determination of object state variables by use of an EKF. The EKF treats the problem of determining the state variables as an estimation process. State variables are first predicted by means of the dynamical model. This prediction is enhanced by the following innovation step. Here, edge elements stemming from edge detection performed in the image are regarded as measurements for points on the object contour. After calculation of the prediction errors these values are arranged in the measurement vector which is used for performing the innovation. While the EKF can be shown to be stable and robust, it does not resolve the matching problem which concerns the assignment of edge elements to objects. As a result, especially in noisy and multi-object environments the EKF is likely to fail. Therefore, the approach presented here uses the GHT to select edge elements grouped in such a way that they are likely to be a part of the projection of an object.

3 Face and facial feature models

As pointed out in the last section, facial features are defined by 3D points in a separate coordinate system for each feature. These points are linked by lines. Fig. 1a gives an example of a simple model of boundaries defined by open eyelids.

Several facial features are arranged in the face model by defining the position and attitude of the facial feature coordinate systems relative to the face coordinate system, cf. Fig. 1b.

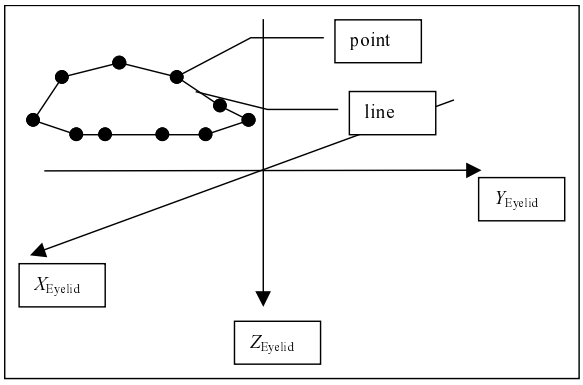


Figure 1a: 3D-model of boundaries with eyelids

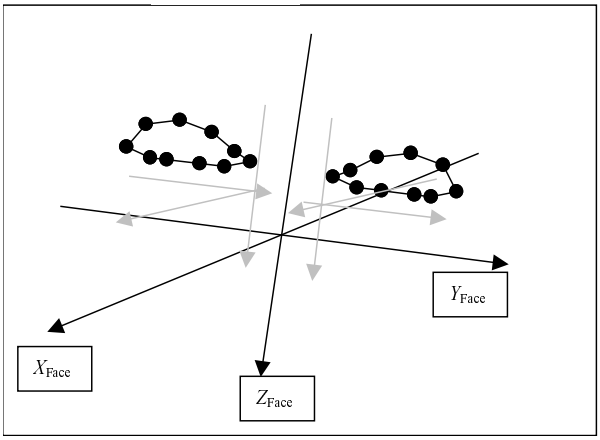


Figure 1b: facial features in face coordinate system

4 General outline of the adaptation of the GHT

Here, the purpose of the adaptation of the GHT is the detection of an object in up to six degrees of freedom (3 translational plus 3 rotational). Depending on the problem at hand a six-dimensional parameter space results. As the GHT is inherently based on discrete values a bandwidth and quantization for each degree of freedom has to be chosen appropriately. As mentioned before, to these six degrees of freedom further parameters determining the shape of the object (shape parameters) may be added.

For the general outline it is assumed that the object representation has been transformed into the camera coordinate system in a nominal relative position and attitude, and the following considerations are to be taken into account. A movement of the parallel to the image plane object (state variables Y and Z) does not entail a change in the projected shape of the object. Under the conditions of the so called weak projection, a movement along the optical axis (state variable X) changes the size but not the shape of the projection like a rotation around the X-axis (state variable Φ) which only changes it's angle relative to the image coordinate system. With respect to changes in the latter two state variables (X and Φ), the resulting shapes can therefore be computed efficiently. However, the projected shape is changed by rotations around the Y and Z- axis. Only in case of variations in the last two state variables (Θ and Ψ, respectively) a new projection into the image plane is indispensable. Translational movement along Y and Z axis as well as a rotational movement around the X -axis can be readily handled by the GHT while for translational movement along the X-axis well known variations of the GHT can be employed.

The adaptation of the GHT also has to correspond to the method by which edge elements are extracted from the image. Here, to cope with real time conditions, fast and efficient correlation with ternary masks (consisting only of –1, 0 and 1 elements, e.g. [5]) are applied along one dimensional search paths which are aligned equidistantly parallel to both axes of the image coordinate system yielding a search grid (fig. 2). These search paths might also be regarded as a sample from the image or as a form of subsampling. In order to determine the form to be detected by the GHT corresponding to this edge detection method first the resulting projection for a given parameter combination of the relevant parameter space has to be computed. For this projection the intersections of the contour with the search path grid are calculated (fig. 2) yielding a derived form spanning a four dimensional space.

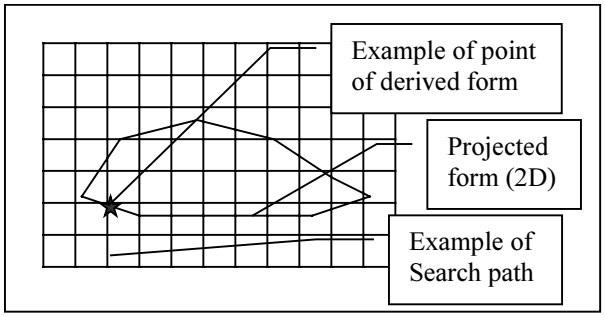


Figure 2: Projected (original) form, search grid and determination of derived form.

Thus, for each derived form, the maximum number of expected edge elements can be calculated and after accumulation a measure of confidence can be calculated by comparison with values of the accumulator arrays which may be used by higher levels of the overall architecture, e.g. [15], [18].

5 Reduction of computational complexity

Reduction of computational complexity can predominantly be achieved by the use of spatio temporal models but also by parameterization of the GHT, i.e. appropriate choice of quantization and bandwidth for the state variables. In order to illustrate the further procedure the following phases which define conditions, under which the detection and measurement task is performed, are distinguished (table 1). Assuming an object consisting of several subobjects, as is the case with faces containing facial features and regarding them in a first approximation as points, a limited number of degrees of freedom (dof) remains, if subobjects are detected (table 1).

Phase (recognition state)	Signification	Degrees of freedom remaining assuming point approximation for subobject	Restrictions
Detection I	No subobject detected	6 dof	None
Detection II	One subobject detected	3 dof (1 translational, 2 rotational)	Detected subobject must lie on a line
Detection III	Two subobjects detected	1-2 dof (1 rotational, 1 translational, restricted)	Each subobject must lie on one of two lines; distance restricted; ambiguity in one rotational dof
Tracking I Kalman Filtering I	All subobjects detected	0 dof	Object position and attitude fully determined
Tracking II Kalman Filtering II	All subobjects detected and gray scale levels along contour known	0 dof	Object position and attitude fully determined

Table 1: Reduction of parameter space

It should be noted that the concept of subobjects can be employed even if such distinct subobjects are not evidently given, since any object may be appropriately subdivided thus enabling complexity reduction. As may be seen from table 1 a dramatic reduction in complexity takes place if an object is partially detected and if geometrical and dynamical models are employed. Evidently, the assumed point approximation for the subobjects is inappropriate since the GHT already yields an estimate for position and attitude of the subobject, thus further reducing complexity. The accuracy of this estimate depends on the quantisation of the GHT. The complexity reduction concerns not only the parameter space but also the image space. Projection of the regions in space in which the so far undetected subobjects are expected reduces considerably the area in which edge detection has to be performed. These regions project (approximately) as circles or parts of circles (Detection II) or as ellipses or parts thereof (Detection III) into the image plane. Since the main objective is reduction of work load in the image plane, exact determination of the regions is not necessary. Rather, they can be approximated by bounding boxes. As the prediction yielded by the Kalman Filter is uncertain, even during the tracking phases, variations in the state variables have to be allowed for. As a result neither the search space in the image nor the parameter space is empty in this case.

During the tracking phases the GHT assists the EKF by selecting the edge elements used by the EKF as measurements. This is especially important in view of the tremendous effects that outliers have on regression analysis, e.g. [13]. The condition "Tracking II" is included in table 1 only for the sake of completeness and will be explained in the next section.

In order to manage the different methods for detection and tracking, the introduction of recognition states for each object and subobject and their modeling as a finite state machine has proven successful (see. [18] for the first application to face detection and tracking and [15] for the further development for hierarchical objects). Broadly speaking, in a video sequence an object starts in recognition state Detection I and eventually moves to the tracking phases in later pictures if subobjects are detected.

Apart from reducing complexity with respect to image and parameter space, this approach minimizes also the possibility of false detection in irrelevant image regions.

6 Improvement of reliability and robustness of GHT

In noisy images, due to a large number of edge elements extracted from the image, false detection may occur. To counter this problem, prior knowledge about the grayscale levels along the contour may be used during the accumulation process of the GHT for selection of edge elements. Two possibilities are presented here:

1. Use of sign of the expected gradient difference on both sides of the contour: this approach is appropriate if no absolute grayscale levels are known but only the fact that the area on one side of the contour is expected to be darker/lighter than the area on the other side. It may therefore be employed during the detection phases I-III and tracking I.
2. If grayscale levels are known along the contour, these may be used for feature selection of edge elements. Since no photometric models are used, the relevant grayscale levels have first to be determined from the image. Therefore, this possibility is used only during tracking phase II.

If in the first case, the sign of the expected grayscale difference for the points of the derived form used by the GHT (section 4) are known, accumulation only takes place if the sign of grayscale level differences of an edge element extracted from the image correspond to the expected one. The intensity direction (dark to light or light to dark) in the 2D-image can be determined by assigning an intensity direction vector to the lines of the 3D-model and projection of this vector into the image plane. All points of the derived form resulting from an intersection of the projected line with the search paths are assigned the same intensity vector.

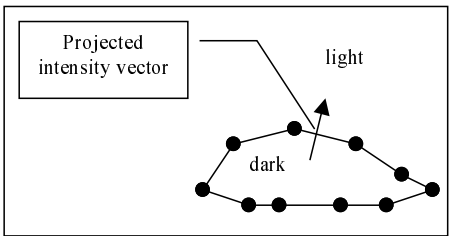


figure 3: Use of sign of intensity gradient for GHT.

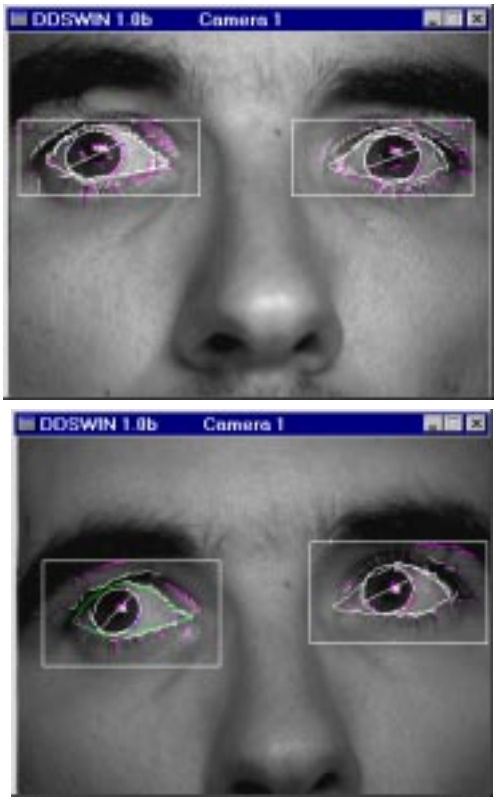
Assuming in the second case that the expected grayscale levels are known for the points of the derived form, accumulation only takes place if the grayscale level(s) of an edge element differ within a threshold limit from the expected level(s) on either or both sides of the contour. In this case, the selection process is much more effective. Since no photometric model is used, the grayscale levels have to be determined and the assignments from 2D to 3D and vice versa have to be solved. The following solution is proposed.

To determine the grayscale levels, the edge elements contributing to the detected maximum of the accumulator array are searched for. These grayscale levels are assigned to the closest element of the original form. Since several elements of the derived form may correspond with an element of the original one, a weighted mean is calculated. In order to improve performance, a history of grayscale levels covering the results of recent images is constantly updated. If the corresponding variance is too high, the corresponding point is not used for accumulation. If it is sufficiently low, the moving average is used for selection. Back assignment of grayscale levels from the 3D original form to the 2D derived form is achieved by linear interpolation between two points of the original form.

The latter version of the GHT is very reliable even in very noisy pictures. Since it uses grayscale levels extracted in previous pictures, it shows a familiarity to pattern matching. Still, it offers advantages. Even though adapting to changing lighting conditions it maintains the invariant 3D structure and can thus cope correctly with 3D movement. Broadly speaking, short term stability is assured by use of grayscale levels and longterm stability by use of the underlying 3D-model. Finally, by monitoring variance, it is able to select only reliable features.

7 Experimental results

The algorithm runs on a Dual Pentium II with 333 MHz in real time at a frame rate of 25 Hz. The experiments have so far been performed in the tele camera because of the required high accuracy using a model of two open eyelids and two iris (cf. [15] for the latter). First, one search window scans the whole image for one eyelid using a wide range of possible aspect conditions. Scanning of the whole image takes about half a second. As soon as one eyelid has been detected, a second search grid starts to scan the image region where the other eye is to be expected according to the 3D-model (cf. section 6). Detection of the second eye starts the Extended Kalman Filter and tracking is performed in the sequel in real time. As mentioned above, first a version of the GHT that uses no intensity information is employed. As soon as enough intensity information has been gathered the version that takes intensity information into account is started. While the first version of the GHT occasionally leads to false detection close to the eye due to the high number of edge elements extracted (eyebrows, eyelashes), the second version that uses short term adaptable intensity information while maintaining the 3D structure is extremely stable even under varying lighting conditions. Fig. 4 gives an example of the algorithm during the tracking phase II. It shows the search fields as light boxes, the projection of the eyelid model into the image plane as continuous white line and detected edges as short light lines. Furthermore, the detected iris are depicted (not treated here, please see [15]) since the purpose of the algorithm is the determination of gaze direction.



Figures 4a and b: Experimental results of tracking phase.

8 Conclusion and future work

An algorithm for detecting and tracking facial features and faces in real time by joint use of spatio-temporal models and GHT-techniques has been presented. For all phases one model is used. It has been pointed out how synergies between GHT and spatio-temporal models including the use of EKFs can be efficiently exploited. The algorithm presented is able to detect and track facial features as well as head movement robustly and reliably in real time with a frame rate of 25 Hz.

Nevertheless, future work will first of all include more facial features (nose, mouth etc. complete face model) the use of which has so far not been feasible because of real time conditions. Therefore, exploitation of the image of the wide angle camera (section 1) will become necessary. The use of 3D models enables the transformation and projection into both cameras. Furthermore, by parameterisation, facial features can be modeled elastically thus allowing for e.g. closure of eyelids which can be incorporated into the 3D-model and the whole algorithm. One of the most important tasks will be the adaptation of a generic model to an individual person. Experiments with the existing algorithm have shown that detection of GHT still works reliably even if different people are to be detected and tracked using one model but errors in the determination of position results. Parameterised models of facial features and 3D reconstruction methods will be used to overcome these difficulties. Since the models employed here are in no other respect special to faces and facial features than their 3D form, the use of the algorithm for detecting and tracking other objects will be examined.

9 References

[1] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*. 13(2), 1981; pp 111-122.

[2] M. J. Black and Y Yacoob. Tracking and recognizing rigid and non rigid facial motions using Local Parametric Models of Image Motion. *Proceedings of the International Conference on Computer Vision*. 1995; pp. 374-381.

[3] D. Dickmanns. *Rahmensystem für visuelle Wahrnehmung veränderlicher Szenen durch Computer*. PhD-thesis Universität der Bundeswehr München Fachbereich Informatik. Neubiberg. 1997.

[4] E. D. Dickmanns, B. Mysliwetz und T. Christians. An Integrated Spatio-Temporal Approach to Automatic Visual Guidance of Autonomous Vehicles. *IEEE Transactions on Systems, Man and Cybernetics*, 20 (6), 1990.

[5] E. D. Dickmanns, V. Graefe. a) Dynamic molocular machine vision. b) Applications of dynamic molocular machine vision. *J. Machine Vision & Application*, November 1988, pp. 223-261

[6] C. Fargerer, D. Dickmanns and E. D. Dickmanns. Visual Grasping with Long Delay Time of a Free Floating Object in Orbit. *Autonomous Robots* (1), 1994, pp. 53-68.

[7] A. J. Glenstrup and T. Engell-Nielsen. *Eye Controlled Media: Present and Future State*. Diploma thesis, Kopenhagen, http://www.diku.dk/~panic/eyegaze/article.html, 1995.

[8] P. V. C. Hough. Method and means for recognising complex patterns. U.S. Patent 3,069,654, 1962.

[9] T. S. Jebara and A. Pentland. Parametrized Structure from Motion for 3D Adaptive Feedback Tracking Faces. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997, pp. 144-150.

[10] H. Klingspohr, T. Block and R.-R. Grigat. Ein echtzeitfähiges System zur Erkennung der Blickrichtung des menschlichen Auges. *Informatik aktuell*, 1997, pp. 191-198.

[11] A. A. Kassim, T. Tan and K. H. Tan. A comparative study of efficient generalised Hough transform techniques. *Image and Vision Computing* (17), 1999, pp. 737-748.

[12] R. Onken and Strohal. The Crew Assistant for Military Aircraft. *Proc. 7th Int. Conf. on Human-Computer Interaction*. San Francisco. 1997.

[13] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. New York, 1987.

[14] A. Schubert and D. Dickmanns. „Real-Time Gaze Observation for Tracking Human Control of Attention. in: H. Wechsler, P. J. Phillips, V. Bruce, F. F, Soulié, T. Huang (eds.). *Face Recognition: From theory to Applications*, Springer, Berlin, 1998, pp. 617-626.

[15] A. Schubert and D. Dickmanns. Bildverarbeitungsalgorithmus zur Vermessung der 3D-Kopfposition und der Blickrichtung eines Menschen. *Proceedings 21. Symposium für Mustererkennung*, Bonn, 1999.

[16] T. Silberberg, D. Harwood and L. Davis. Object Recognition Using Oriented Model Points. *Computer Vision, Graphics, and Image Processing* (35) 1986, pp. 47-71.

[17] K. Sobottka and Ioannis Pitas. Segmentation and Tracking of faces in Color Images. *Proceedings of the second international conference on face and gesture recognition*, 1996, pp. 236-241.

[18] K. Toyama and G. D. Hager. Increment Focus of Attention for Robust Visual Tracking. *Proc. Computer Vision and Pattern Recognition*, 1996, pp. 940-945.

[19] H. Wechsler, P. J. Phillips, V. Bruce, F. F, Soulié and T. Huang (eds.). *Face Recognition: From theory to Applications*, Springer, Berlin, 1998.