Eye Localization Using The Discriminative Generalized Hough Transform

Ferdinand Hahmann[1], Heike Ruppertshofen[2], Gordon Böer[1], Ralf Stannarius[3], Hauke Schramm[1]

[1]University of Applied Sciences Kiel, [2]Philips Technologie GmbH, [3]Rosemann Software GmbH

{Ferdinand.Hahmann,Gordon.Boeer,Hauke.Schramm}@fh-kiel.de, Heike.Ruppertshofen@philips.com, Ralf.Stannarius@rosemann-software.de

Abstract. The Discriminative Generalized Hough Transform (DGHT) has been successfully introduced as a general method for the localization of arbitrary objects with well-defined shape in medical images. In this contribution, the framework is, for the first time, applied to the localization of eyes in a public face database. Based on a set of training images with annotated target points, the training procedure combines the Hough space votes of individual shape model points into a probability distribution of the maximum-entropy family and optimizes the free parameters of this distribution with respect to the training error rate. This assigns individual positive and negative weights to the shape model points, reflecting important structures of the target object and confusable shapes, respectively. Additionally, the estimated weights allow to determine irrelevant parts in order to eliminate them from the model, making space for the incorporation of new model point candidates. These candidates are in turn identified from training images with remaining high localization error. The whole procedure of weight estimation, point elimination, testing on training images and incorporation of new model point hypotheses is iterated several times until a stopping criterion is met. The method is further enhanced by applying a multi-level approach, in which the searched region is reduced in 6 zooming steps, using individually trained shape models on each level. An evaluation on the PUT face database has shown that the system achieves a state-of-the-art success rate of 99% for iris detection in frontal-view images and 95% if the test set contains the full head pose variability.

1 Introduction

The localization of facial features like the eye is an important prerequisite for a variety of applications in the field of computer vision [3]. Because of its considerable importance, this topic attracted a lot of attention by the scientific community in the past which is reflected in a large variety of different approaches. It is remarkable, that most of the current state-of-the-art systems rely on the usage of a face detector before applying specifically adapted methods for eye localization. The most popular face detection method has been proposed in [17].

This technique uses Haar-Wavelet-Features in a boosted cascade of classifiers to identify regions containing a face. Since the general method can be trained to detect arbitrary objects with well-defined appearance, this face detector has been combined in [12] with a subsequent eye localization, based on the same technique. Although the first step of this procedure substantially reduced the search space, additional knowledge about the specific task had to be incorporated to further restrict the number of hypotheses of the classifier cascade. The usage of task-specific knowledge is prevalent in many state-of-the-art systems published in recent years. In [13], for example, the approximate location of the eye in relation to a given face bounding box is determined and the estimation is refined by searching for a minimum intensity region in a defined neighborhood, which is expected to be the pupil. In [16] the search for a circular, dark structure is performed by means of the orientation of all image gradients, weighted by the gray value, to determine the centers of dark circles (like the pupil).

Further systems determine location hypotheses for both eyes and reject those not meeting the expected eye distance [6] or the allowed in-plane rotation [12]. Only few methods, however, learn these constraints from data, like in [5], where 15 additional facial features are detected together with the eyes and combined with learned feature distance probability distributions.

The usage of expert knowledge can substantially improve the performance of an object localization framework but requires in-depth knowledge about both the task and the system. Yet, such solutions can usually not be generalized to other problems. In addition, the combination of different specialized methodologies into a single framework may lead to complex systems, which are difficult to maintain and advance.

A general and widely used approach of object localization is the Generalized Hough Transform (GHT)[1]. Many extensions to the GHT have been developed in recent years to achieve a reduced processing time and increased robustness, like the Hough Forests [7] or the Discriminative Generalized Hough Transform (DGHT) [15], which has been applied in this work. The DGHT is a general object localization method, initially developed in the field of medical image processing [15], which aims at automatically generating optimal shape models for usage in a standard Generalized Hough Transform. It estimates individual positive and negative model point weights for a given shape model and eliminates irrelevant parts in a subsequent step. In an iterative training procedure, this technique is applied repeatedly to expand the model with image structures of unrecognized targets and confusing objects in the training corpus. In this work, the DGHT method will, for the first time, be applied to an eye localization task. The procedure operates without incorporating additional methods or expert knowledge and achieves state-of-the-art localization rates on a given public database.

Here, section 2 describes the general method, including the GHT (2.1), the discriminative training of GHT models (2.2), and a zooming strategy to improve the robustness and processing time (2.3). Section 3 contains a description of the used database (3.1), the experimental setup (3.2), and lists the achieved results (3.3) followed by a discussion (3.4). Section 4 concludes our contribution.

2 Method

This section presents the Discriminative Generalized Hough Transform (DGHT), a method recently developed in the field of medical image analysis [15]. The technique combines the well-known Generalized Hough Transform (GHT) [1] with a discriminative training technique [2] to generate sparse, efficient and discriminative GHT-models for object detection.

2.1 GHT

The GHT was introduced by Ballard in 1981 [1]. It is a general and well-known model-based method for object localization, widely used for the analysis of 2-D images.

The GHT requires a point model M, representing the structure of the searched-for object in relation to a so-called reference point. This point, usually located in the center of the model, represents the target point of the object detection task in the DGHT framework.

By performing a simple voting procedure, the GHT transforms a feature image X_n into a parameter space H, known as the Hough space. This space consists of accumulator cells, representing possible target point locations and, potentially, model transformations. The amount of votes in an individual cell reflects the degree of matching between the (possibly transformed) GHT model and the feature image. Note that, in this work, no model transformations apart from translations are considered since moderate object variability with respect to shape, rotation and size is learned into the model. In addition, although arbitrary features could be applied in the general GHT framework, the presented system relies on plain edge features, obtained by a standard Canny edge detection [4].

The GHT voting procedure considers all possible combinations of edge points $\mathbf{e}_i$ and model points $\mathbf{m}_j$ and increments the accumulator cell at the image coordinates $\mathbf{x}$ as follows:

$$H(\mathbf{x}) = \sum_{\forall \mathbf{e}_i \in X_n} \sum_{\forall \mathbf{m}_j \in M} \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{e}_i - \mathbf{m}_j \text{ and } \arccos(< \varphi_i, \varphi_j >) \leq \vartheta_\varphi \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Here, φ_i and φ_j denote the gradient direction of $\mathbf{e}_i$ and orientation of $\mathbf{m}_j$, respectively. A point pair is allowed to vote if it has a similar direction, meaning that the scalar product must be below the threshold ϑ_φ. In order to speed-up the GHT voting procedure, Ballard introduced the so-called R-table [1]. In this look-up table, each bin is assigned a certain range of angels $\Delta\Phi$ and model points are sorted into the bins according to their orientation. During the voting procedure, an edge point $\mathbf{e}_i$ only votes with the model points contained in the bin with the corresponding orientation range.

Obviously, the performance of the GHT-based localization procedure heavily relies on the choice of an appropriate model. In addition to the geometric distribution of the applied point cloud, this may also include a sophisticated weighting scheme of individual points. This is quite reasonable since it can be

expected that some parts of a given model are more important than others and should therefore vote with a higher weight. In [15] a novel approach for a discriminative training of GHT model point weights has been presented, including the estimation of both positive and negative weights.

2.2 Discriminative Training of GHT Models

The DGHT framework applies an iterative training procedure [15], which determines the geometric model layout as well as individual weights of the model points in order to take the different importance of model structures into account. To this end, a subset of images from the training corpus with annotated target landmarks is applied to generate an initial GHT model. This is achieved by using the Canny edge points from the selected images as model points and the gradient directions as their orientation. A subsequently performed individual weighting of the model points is achieved by the discriminative training procedure described below. Since model points with a low absolute weight have only a small influence on the resulting Hough space, they are discarded from the model in order to make space for new model point candidates. After that, the model is tested on the whole annotated training dataset to detect images with remaining high localization error. These images are used to learn unknown target object shapes and important rivaling objects which are both incorporated into the model by using the respective edge points as additional model points for the next iteration. The whole procedure, which is described in detail in [15], is iterated until a given stop-criterion is met.

The estimation of individual weights for the J points, contained in a given shape model, is achieved by (1) separating the Hough space votes coming from every single model point, (2) recombining those contributions in a weighted manner and (3) optimizing the introduced weights with respect to an error measure. The theory of this approach is based on describing the GHT as a probabilistic framework in which the Hough space is interpreted as a posterior probability distribution $p(c_i|X_n)$, estimated from the relative frequencies of the model point votes in each Hough cell c_i. The object localization task, performed by the GHT, which searches for the cell with the highest number of votes, can now be formulated as the Bayes classifier $\hat{c} = \arg\max_{c_i} p(c_i|X_n)$. In order to split the total number of votes in a Hough cell c_i into contributions coming from individual model points, we introduce the feature function

$$f_j(c_i, X_n) = v_{i,j}, \tag{2}$$

which denotes the number of votes $v_{i,j}$ from model point m_j into cell c_i for a given image X_n. A weighted recombination of those features in a maximum entropy distribution [8]

$$p_\Lambda(c_i|X_n) = \frac{\exp\left(\sum_j \lambda_j \cdot f_j(c_i, X_n)\right)}{\sum_k \exp\left(\sum_j \lambda_j \cdot f_j(c_k, X_n)\right)} \tag{3}$$

preserves objectivity, incorporates the constraints from the GHT voting procedure with the given model and introduces model point specific weights λ_j. The estimation of those free parameters from the side conditions (2) leads to an optimal approximation of the training data distribution. Since this does not necessarily lead to a minimal error rate, we follow an alternative approach, first presented in the field of automatic speech recognition [2]. In our framework, the model point weights λ_j are optimized by using a Minimum Classification Error (MCE) training approach [10], which minimizes a smoothed error measure over the N training images and I Hough cells:

$$E(\Lambda) = \sum_{n=1}^{N} \sum_{i=1}^{I} \varepsilon(c_i, \widetilde{c}_n) \cdot \frac{p_\Lambda(c_i|X_n)^\eta}{\sum_k p_\Lambda(c_k|X_n)^\eta}. \tag{4}$$

Here, η in the smoothing term controls the influence of alternative location hypotheses on the error measure and $\varepsilon(c_i, \widetilde{c}_n)$ denotes the localization error between the Hough cell c_i and the target Hough cell $\widetilde{c}_n$ using the Euclidean distance:

$$\varepsilon(c_i, \widetilde{c}_n) = \|c_i, \widetilde{c}_n\|_2 . \tag{5}$$

In the DGHT framework, the optimization of $E(\Lambda)$ over the model point weights $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_J\}$ is achieved by applying the method of steepest descent. Although this technique does not assure to find a global minimum, recent experiments in the field of medical image analysis have shown substantial improvement compared to equal weighting strategies or other state-of-the-art methods [15].

The estimated weights of the shape model are directly incorporated into a standard GHT voting procedure by incrementing the value of a Hough cell c_i by $\lambda_j \cdot f_j(c_i, X_n)$ for each model point m_j:

$$H(c_i|X_n) = \sum_{j=1}^{J} \lambda_j \cdot f_j(c_i, X_n). \tag{6}$$

The localization result, i.e., the detected Hough cell $\hat{c}$, is then given by $\hat{c} = \arg\max_{c_i} H(c_i|X_n)$. Note that the usage of this weighting scheme in the localization leads to the same result as applying the maximum entropy distribution (3) used for the optimization. This is, since neither the normalization term in the denominator of (3) nor the exponential function have an influence on the result of the arg max function.

2.3 Multi-Level-Approach

In order to speed-up the localization procedure and to increase the robustness of the system, we follow a coarse-to-fine strategy, as described in [14]. A Gaussian image pyramid is created and the localization is performed on each level with a specifically trained shape model. The first resolution level is very coarse and shows little detail but allows for a rough orientation and fast processing since the

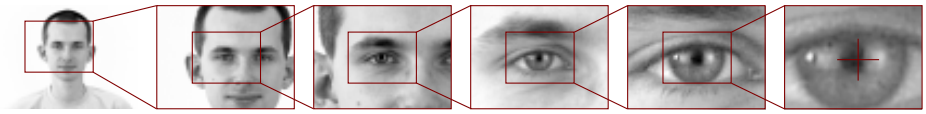

Fig. 1: Image extracts with different resolutions in the multi-level approach.

image size is small and the shape model contains only a restricted set of model points with little object detail. The detected point, found with this model, is then used to determine the center of the image extract, used for the next level (see Fig. 1). This extract has half of the size of the original image and a higher resolution than the previous level. The procedure can be repeated several times, each time using a specifically trained and more detailed DGHT model.

Since the image extracts have only a fraction of the original image size, they can be processed quickly, even with the full resolution. In addition, the shape model used in the final level does not need to represent global details since these have already been captured by the previous levels. This specialization of the individual shape models to level-specific details and characteristics aims at reducing the localization error by finding an optimal trade-off between sufficient model accuracy and reduced confusion with concurrent objects.

3 Experiments

3.1 Data

The training and evaluation set used in the experiments is extracted from the publicly available PUT face database [11], which contains 9971 images from 100 subjects. The color images were taken under controlled illumination conditions with a uniform background and have a high resolution of 2048×1536 pixels. Manual annotations are provided for various landmarks, e.g., in the center and along the outline of the eyes, nose, face, eyebrows and mouth. Note that those landmarks were only annotated for images where the corresponding facial feature is visible. Despite the uniform background, this database is demanding since it covers a large variation of head poses as shown in Fig. 2, making it an interesting base for localization experiments. In order to study the influence of head pose variations on the localization error rate, a frontal view subset of the data, containing 22 images per subject, was provided as well.

The subjects, contained in the PUT database, were strictly separated during training and evaluation. 60 randomly selected subjects were used for training and the remaining 40 subjects for evaluation. In the experiments reported here, the target point for the localization was always the center of the eye. Since, due to the strong head pose variation, this landmark is not visible in all the images of the database, only a subset with given annotation of this point was used. The training database contains 10 randomly selected images per subject, in total 600 images. The evaluation was done on 3830 images from the 40 test subjects. The frontal view evaluation subset contains 869 images.

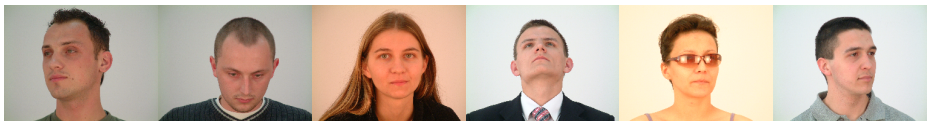

Fig. 2: Illustration of the large head pose variability, contained in the PUT database.

Table 1: Success rates for DGHT eye localization on the evaluation corpus using error measure (7) and different fault tolerances.

	$\mathbf{e < 0.1}$	$\mathbf{e < 0.15}$	$\mathbf{e < 0.2}$	$\mathbf{e < 0.25}$
frontal view subset	99.08%	99.08%	99.31%	99.42%
full head pose variations	95.01%	95.43%	96.01%	96.48%

3.2 System setup

In all experiments, we used Canny edge detection [4] for the generation of feature images and applied an iterative training [15] and multi-level approach (Section 2.3). In the multi-level framework six zooming levels were used, ranging from the whole image to a small region around the pupil as shown in Fig. 1. In each zooming level, the resolution is 64×48 pixels. The experiments were performed on a 64 bit system with an Intel Xeon W3520 with 2.66 GHz and 24 GB RAM.

To validate the localization success, we use a relative error measure, which was introduced in [9]. It selects the worst of a left and right eye localization and normalizes this error distance to the respective eyes distance:

$$e = \frac{max(||\mathbf{c}_l - \widetilde{\mathbf{c}}_l||, ||\mathbf{c}_r - \widetilde{\mathbf{c}}_r||)}{||\widetilde{\mathbf{c}}_l - \widetilde{\mathbf{c}}_r||}. \tag{7}$$

Here, the vectors $\mathbf{c}_l$ and $\mathbf{c}_r$ are used to denote the localization result for left and right eye localization while $\widetilde{\mathbf{c}}_l$ and $\widetilde{\mathbf{c}}_r$ are the annotated left and right eye positions. Note that, using this measure, an error of less than 0.1 / 0.25 relates to a localization result lying approximately within the iris / eye.

3.3 Results

The iterative training performed 24 iterations on average for the different zooming levels and achieved a final error on training data of 4%. The average number of points in the level-specific shape models is 357, ranging from 180 to 395 points.

The rates for a successful iris localization ($e < 0.1$), shown in Table 1, are competitive with the results reported in [12], where success rates of 99% and 94% were achieved on this database for frontal views and full head pose variations, respectively.

The average processing time measured for a single image was 597 ms, while the GHT voting procedure and identification of the best Hough cell hypothesis took only 1% of the complete runtime.

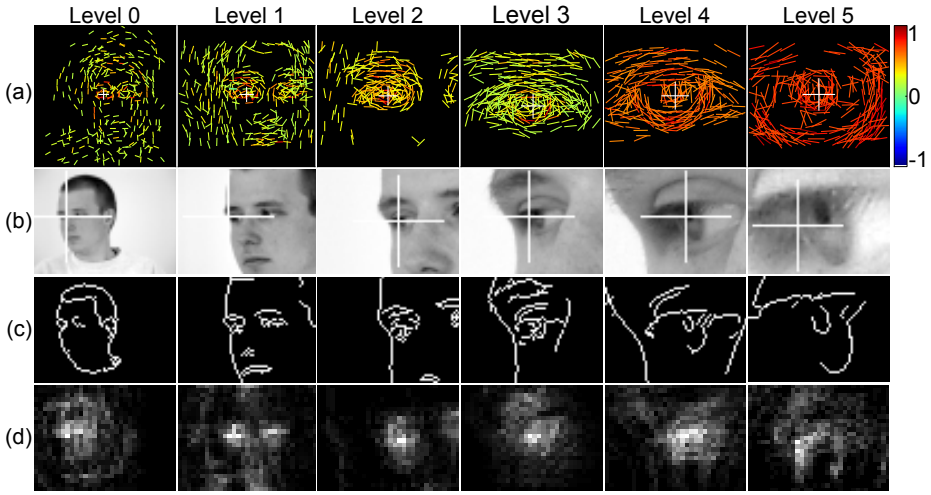

Fig. 3: (a) shows the models for all six zooming levels, (b) - (d) shows a complete example including the result image (b), the feature image (c) and the Hough space (d).

3.4 Discussion

Fig. 3(a) presents the models of all six zoom levels for a right eye localization. The model points are represented as lines, to visualize their orientation. It can be seen that the level of detail increases substantially from the global to the single eye model, similar to the visible structures in the feature images. At level 0, the model contains information about different head sizes and hairstyles, while the eyes and mouth are only coarsely represented. Since global structures like hairs are no longer visible in higher levels, the models may focus on the final target, which leads to a highly detailed iris and pupil modeling.

It is noticeable that in levels 0 to 3 significant differences between the model point weights are visible while the weights in levels 4 and 5 are similar and quite high. This is due to the slightly inaccurate target point annotations (shown in Fig. 4), which, naturally, have the strongest effect on the smallest image extracts. Therefore, the stop criterion in the training procedure was met only after performing a large number of iterations, causing the observed high weights. The random character of these annotation errors also prevents the discrimination of important and irrelevant model points and leads to the more or less equal weighting scheme, visible in Fig. 3.

In the edge image of level 5, it is, furthermore, visible that, in addition to the pupil, also the flash light must be represented by the model. This can indeed be observed on the left and right of the models' pupil structure.

An error analysis unveiled that 61% of the false localizations can be accounted to a confusion of the eyes. In these cases, the independently applied left and right eye detector returned the same result. To this end, it can be assumed that the



Fig. 4: Illustration of manual annotations, provided by the PUT face database

integration of constraints, like, e.g., a learned eye distance, will result in a higher localization performance.

It should also be mentioned that first investigations have confirmed, that a substantial runtime improvement can be expected from using faster implementations for the downsampling and feature extraction, since the employed slow procedures require most of the computing performance.

4 Conclusion

In this contribution, we introduced the usage of the Discriminative Generalized Hough Transform (DGHT) for eye localization. The method, which was originally developed in the field of 2-D and 3-D medical image processing, is a general localization technique for arbitrary objects with well-defined shape. Since the iterative training method simply requires a set of images with annotated target points and runs fully automatic without the need of expert knowledge, it can be quickly applied to new tasks.

It should be mentioned that, different to most other state-of-the-art eye localization techniques, no additional method for face localization is required. The first rough orientation step is acquired by the same general approach, simply using a specifically trained model. First experimental results in this new application field show that the DGHT can be successfully applied to eye localization in a controlled environment with a uniform background. Despite substantial head pose variations in the PUT face database, a localization rate of 95% was achieved for iris detection, which is competitive to other methods recently tested on this dataset [12]. This result improves to 99% for a subset of the data containing frontal view images only. An interesting aspect of the presented framework is the possibility of a straightforward interpretation of the resulting models due to their weighted geometric structure. It can be seen that the contour of the head and eyes is of special importance for the coarse localization of the face while the model of the latest zooming level has even learned lighting conditions reflected in the eye. Additionally, it should be stressed that the method operates without incorporation of any task specific constraints, defined, e.g., by an expert. It can, however, be expected that further gains in the success rate could be achieved by exploiting additional knowledge, for example about the expected distance between the target object and other landmarks in the face. The incorporation of this idea into the given framework is quite straightforward and will be investigated in future work.

Acknowledgments. This work is partly funded by the Innovation Foundation Schleswig-Holstein under the grant 2010-90H .

References

1. Ballard, D.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Recognition 13(2), 111–122 (1981)
2. Beyerlein, P.: Discriminative model combination. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 481–484 (1998)
3. Böhme, M., Meyer, A., Martinetz, T., Barth, E.: Remote eye tracking: State of the art and directions for future development. In: Conference on Communication by Gaze Interaction (COGAIN). pp. 12–17 (2006)
4. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
5. Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: British Machine Vision Conference (BMVC). pp. 277–286 (2004)
6. D'Orazio, T., Leo, M., Cicirelli, G., Distante, A.: An algorithm for real time eye detection in face images. In: International Conference on Pattern Recognition (ICPR). pp. 278–281 (2004)
7. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1022–1029 (2009)
8. Jaynes, E.: Information theory and statistical mechanics. The Physical review 106(4), 620–630 (1957)
9. Jesorsky, O., Kirchberg, K., Frischholz, R.: Robust face detection using the hausdorff distance. In: International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA). pp. 90–95 (2001)
10. Juang, B., Katagiri, S.: Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing 40(12), 3043–3054 (1992)
11. Kasinski, A., Florek, A., Schmidt, A.: The PUT face database. Image Processing and Communications 13(3-4), 59–64 (2008)
12. Kasinski, A., Schmidt, A.: The architecture and performance of the face and eyes detection system based on the haar cascade classifiers. Pattern Analysis & Applications 13(2), 197–211 (2010)
13. Kroon, B., Hanjalic, A., Maas, S.: Eye localization for face matching: is it always useful and under what conditions? In: International Conference on Content-based image and video retrieval (CIVR). pp. 379–388 (2008)
14. Ruppertshofen, H., Künne, D., Lorenz, C., Schmidt, S., Beyerlein, P., Salah, Z., Rose, G., Schramm, H.: Multi-level approach for the discriminative generalized hough transform. In: Computer- und Roboterassistierte Chirugie (CURAC). pp. 67–70 (2011)
15. Ruppertshofen, H., Lorenz, C., Schmidt, S., Beyerlein, P., Salah, Z., Rose, G., Schramm, H.: Discriminative generalized hough transform for localization of joints in the lower extremities. Computer Science-Research and Development 26(1), 97–105 (2011)
16. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. In: Conference on Computer Vision Theory and Applications (VISAPP) (2011)
17. Viola, P., Jones, M.: Robust real-time face detection. International journal of computer vision 57(2), 137–154 (2004)