MAM: Transfer learning for fully automatic video annotation and specialized detector creation

Wolfgang Fuhl[1], Nora Castner[1], Lin Zhuang[2], Markus Holzer[2], Wolfgang Rosenstiel[1], and Enkelejda Kasneci[1]

[1] Eberhard Karls University, Sand 14, 72076 Tuebingen, Germany, {fuhl,castnern,rosenstiel,kasneci}@informatik.uni-tuebingen.de
[2] Robert Bosch GmbH, Car Multimedia, Germany, 71272 Renningen, Germany, {Zhuang.Lin,Markus.Holzer}@de.bosch.com

Abstract. Accurate point detection on image data is an important task for many applications, such as in robot perception, scene understanding, gaze point regression in eye tracking, head pose estimation, or object outline estimation. In addition, it can be beneficial for various object detection tasks where minimal bounding boxes are searched and the method can be applied to each corner. We propose a novel self training method, *Multiple Annotation Maturation (MAM)* that enables fully automatic labeling of large amounts of image data. Moreover, MAM produces detectors, which can be used online afterward. We evaluated our algorithm on data from different detection tasks for eye, pupil center (head mounted and remote), and eyelid outline point and compared the performance to the state-of-the-art. The evaluation was done on over 300,000 images, and our method shows outstanding adaptability and robustness. In addition, we contribute a new dataset with more than 16,200 accurate manually-labeled images from the remote eyelid, pupil center, and pupil outline detection. This dataset was recorded in a prototype car interior equipped with all standard tools, posing various challenges to object detection such as reflections, occlusion from steering wheel movement, or large head movements. The data set and library are available for download at *http://ti.uni-tuebingen.de/Projekte.1801.0.html*.

Keywords: Automatic annotation, detector creation, eyelids, eye detection, training set clustering, pupil detection

1 Introduction

Modern applications from diverse fields rely on robust image-based object detection. These fields include, though are not limited to, autonomous driving [8, 4] and scene understanding [36], driver monitoring [5, 28], eye tracking [10, 51], cognitive sciences [54], psychology [29], medicine [16] and many more. To approach object detection, many leading techniques are based on Deep Neural Networks, and in particular, on Convolutional Neural Networks [33, 50]. Recent improvements of CNNs are multi-scale layers [23], deconvolution layers [62]

(transposed convolutions), and recurrent architectures [41, 44]. Nevertheless, the main disadvantage of such networks is that they need an immense amount of annotated data to obtain a robust and general network. For instance, in the realm of eye-tracking, gaze position estimation and eye movement detection are based on robust detection of the pupil center from eye images [20]. More specifically, modern eye trackers rely on image-based pupil center detection and head pose estimation, where multiple landmarks have to be initially detected. A state-of-the-art approach to cope with this problem is to synthesize image data. For example, [48] employed rendered images for gaze position estimation in both head-mounted and remote eye tracking. [32] and [45] used rendering to measure the effect of eyeglasses on the gaze estimation. [59] applied a k- nearest neighbor estimator on rendered images to compute the gaze signal of a person directly from an image. This approach was further improved by [63] using rendered data to train a CNN.

Also, rendering data itself is challenging, since the objective is for highly realistic data that not only cover a certain variety of anatomical structures of the eye and head, but also reflect realistic image capturing properties of the real world. Consequently, models generally need to be trained on both synthetic and real images. Since the annotation of real-world images is a tedious task, we propose an algorithm supporting accurate image annotation: Coined as *Multiple Annotation Maturation (MAM)*. MAM is a self training algorithm based on a grid of detectors. Unlabeled data is clustered based on the detection, iteration, and recognition. To ensure a high detection accuracy for each point, our approach uses a grid of detectors. The deformation of this grid is used to cope with object deformation and occlusions. MAM enables labeling of a large amount of data based only on a small fraction of annotated data and is also capable of reusing already trained detectors under different environmental conditions. Additionally, it delivers specialized object detectors, which can further be used for new data annotations or online detection.

The remaining of this paper is organized as follows. After a review of related work on transfer learning, the proposed approach is described. We show examples of our new dataset as well as how it was annotated. The last sections are the evaluation of the proposed approach on public datasets and its limitations.

2 Related Work

Our method belongs to the domain of transfer learning. Transfer learning itself refers to the problem of adapting a classification model or detector to a new problem, or enhancing its general performance on unknown data. This problem can be solved in an inductive, transductive, or unsupervised way. In the inductive case, annotated data in the target domain is provided in addition to labeled data from the source domain. The process is called self-thought learning or multi-task learning. In self-thought learning, unlabeled data is used to improve the classification performance. For example, [42] proposed a two-step architecture: Where in the first step, feature extraction is improved by analyzing

the unlabeled data using sparse coding [38]. The obtained basis vectors are used afterward to generate a new training set from the labeled data. Then, a machine learning approach, such as a support vector machine (SVM), is trained on the new training data. In multi-task learning, the goal is to improve the classification based on the information gain from other tasks or classes. It has been shown experimentally in [2, 7, 52, 11] that if the tasks are related to each other, multi-task learning outperforms individual task learning. In [2] for example, a Gaussian Mixture Model on a general Bayesian statistics-based approach as developed by [3, 1] was employed. [11] developed a nonlinear kernel function similar to SVMs, which couples the multi-task parameters to a relation between two regularization parameters and separated slack variables per task. In another work, [6] inspected the problem of detecting pedestrians in different datasets, where the recording system differed (DC [35] and NICTA [37]). The authors used a nearest neighbor search to adapt the distribution between the training sets of both data sets to construct a new training set.

In the transductive case of transfer learning, available labeled data in the source domain is employed with the intention to adapt the model to a new (but related) domain, i.e., domain adaption. In this case, the domain is same; however, the problem is reduced to the sample selection bias. Meaning, finding the weighting of training that trains a better-generalized classification, as proposed by [25]. Another approach is the covariance shift proposed by [46], which is the importance weighting of samples in a cross-validation scenario with the same goal of producing a better-generalized classification. If the domain or distribution between the training set and the target set differs, it is usually known as domain adaption. Numerous works have been proposed in this field of transfer learning. For example, [24] proposed Large Scale Detection through Adaptation (LSDA), which learns the difference between the classification and the detection task to transform labeled classification data into detection data by finding a minimal bounding box for the classification label. [43] adapts a recurrent convolutional neuronal network detector (RCNN) trained on labeled data to unlabeled data. Here, the first step is normalizing the data in the source and target domain by calculating the first n principal components. Afterwards, a transformation matrix aligning both domains is computed. The source data is then transformed using this matrix; afterwards, the RCCN detector is trained on the transformed data. For example, in [26], a Gaussian process regression was used to reclassify uncertain detections of a Haar Cascade classifier [56]. The Gaussian process is initialized based on certain detection values that were chosen threshold based. In [12], domain adaption was used to improve image classification. Their proposed pipeline starts with maximum mean discrepancy (same as in [34, 47, 39]) for a dimensionality reduction and aims to minimize the distance of the means of the source and target domain. Afterwards, a transformation based on Gaussian Mixture Models is computed and applied to the source domain. This step aims to adjust the marginal distribution. The last step is a class-conditional distribution adaption as proposed in [13], which is based again on Gaussian Mixture Models. The same procedure is used in [34], where a modified version of the maximum

mean discrepancy is used for the marginal and conditional distribution adaption. [47] learned a nonlinear transformation kernel as first proposed by [40], with the difference being they used eigendecomposition to avoid the need for semidefinite programming (SDP) solvers. In the realm of deformable part-based models, [61] proposed to incrementally improve the target classifier, basing it on multiple instance learning. Therefore, their model needs either some ground truth in the target data or a previously trained detector. For a new image, training data is updated based on the detections and retrain detectors on this data. This step is repeated until there is no update to the training set.

The last (and the most challenging) category is unsupervised learning. The most famous representer of this group is the Principal Component Analysis [58]. The main application of unsupervised learning is the feature extraction (from images or from audio data) [39] based on autoencoders. The signal itself is the target label and the internal weights are learned as a sparse representation. This representation serves as an easier, understandable structure of input data for machine learning algorithms. Based on such features, more advanced approaches like one-shot object classification, as proposed by [14] or one-shot gesture recognition by [57] can be applied. [14] initialized a multi-dimensional Gaussian Mixture Model on already learned object categories and retrained it on a small set of new object classes using Variational Bayesian Expectation Maximization. [60] proposed new feature extractor which is the extended motion history images. It includes gait energy information (compensation for pixels with low or no motion) and the inverse recording (recover the loss of initial frames).

Our approach for automatic video labeling belongs to the category of self-training. It does not require prior knowledge of the object to detect, rather a very small set of labeled examples. It can be done by either using a previously trained detector, or by labeling some object positions (ten in the evaluation).

3 Method

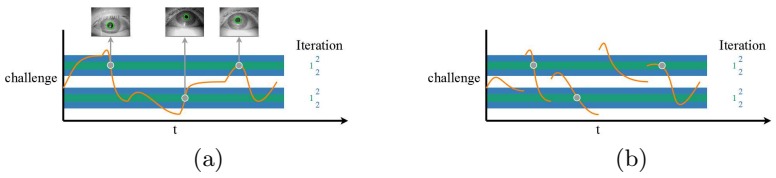


Fig. 1. Our approach, MAM, tries to extend its knowledge of the object. The orange line represents the object to be detected in the video under different conditions such as reflections or changing illumination (challenges). The x-axis represents the timeline of the video, whereas gray dots represent the initially given labels. The green bar represents the detected objects representing similar challenges. Blue is the detection state after the second iteration.

The general idea behind our algorithm is that an object occurs under similar conditions in a video, but at different timestamps. With similar conditions, we mean equal pose and illumination for example. Therefore, different conditions cause varying challenges. As illustrated in Figure 1(a), the orange line represents the same object under different conditions (y-axis) over time (x-axis). Using this interpretation, we can consider the object in a video as a function (orange line). Given some examples (gray dots in Figure 1), our algorithm tries to detect objects under similar conditions in the entire video (horizontal juxtaposed dots on the orange line). The knowledge gain out of the first iteration is represented as the green bars in Figure 1. In the second iteration, this knowledge is extended (blue bars) by retraining a detector on the existing knowledge. This approach alone leads to saturation, which is especially present if some challenges are over-represented in the video. Even more, it can occur if the object does not follow a continuous function, which also impedes tracking (orange line Figure 1(b)).

To cope with this problem, we propose to cluster the detections (knowledge K) into age groups (A, Equation 3); where the age is determined by the amount of re-detections. This clustering allows us to train a set of detectors for different age groups. The detector, which is trained on the entire knowledge obtained from the video (V), is for validation of new detections over multiple iterations (re-detection). The detectors trained on a younger subsets are used to extend the knowledge. Then, the challenge becomes evaluating whether a newly trained detector is reliable or not. Here, we use a threshold TH on recall and precision (on the training set). If both are below TH, the algorithm is stopped or the detector is excluded from the iteration (Equation 1).

$$STOP = \begin{cases} 1 & \frac{TP}{TP+FP} < TH \\ 1 & \frac{TP}{TP+FN} < TH \end{cases} \tag{1}$$

$$\begin{gathered} D^{Age}_{Iter,Feat} = \frac{1}{2}||w||^2 \sum_i^{|A<Age|} \alpha_i \\ (y_i \in L_{A<Age}(\langle x_i \in Feat(K_{A<Age}), w\rangle + b) - 1) \end{gathered} \tag{2}$$

Equation 2 shows the simplified optimization of an SVM for the age subsets (used in this work). w is the orthogonal vector to the hyperplane, α is the Lagrange multipliers, and b is the shift. In this optimization, we seek to maximize α and minimize b, w. With $L_{A<Age}$, we address the subset of found labels L, which has a lower age than Age. The same applies for $K_{A<Age}$, where $Feat()$ represents a transformation of the input data. In our implementations, we only used the raw and histogram equalized images. The detector $D^{Age}_{Iter,Feat}$ can be any machine learning algorithm, e.g. CNN, random forest, neuronal net, etc.

$$A(i) = \begin{cases} A(i) += a & , K(i) \in D^{Age}_{Iter,Feat}(V) \\ 0 & , else \end{cases} \tag{3}$$

Equation 3 specifies the aging function. If the detector D^{Age}_{Iter} detects a previously found object on an image, the age of this object is increased by a constant factor a. In the following, we will describe the details of our algorithm and address our

solutions for the challenge of detecting the position accurately without further information about the object (avoid drifting).

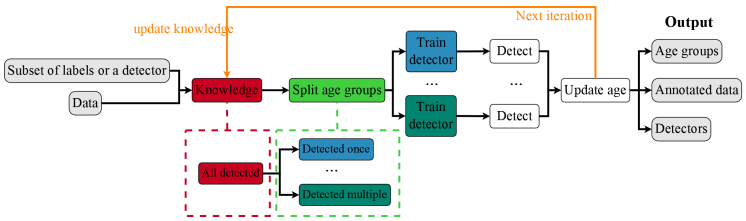
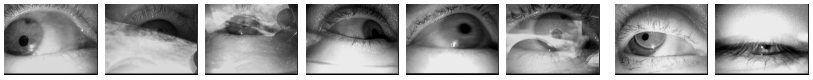


Fig. 2. Workflow of the MAM algorithm. The gray boxes on top represent the input and on the bottom, the output for each iteration. The algorithm starts by splitting its knowledge into age groups and trains detectors for each of them. Afterwards, knowledge and age are updated and a new iteration starts (orange arrow).

Figure 2 shows the workflow of the algorithm, where either a previously labeled set or a detector can serve as input. The input represents the initial knowledge of the algorithm. In the first iteration, only one detector can be trained (since only one age group exists). After n iterations, there can be theoretically n age groups, though this does not happen in practice. Nonetheless, it is useful to restrict the number of age groups for two reasons. First, it reduces the computational costs in the detection part (since each detector has to see the entire video). Second, it packs together similar challenges, which would generate more stable detectors. For all our implementations, we used three age groups. The first group ($G1$) trains on the entire knowledge for validation (Equation 1) and correction. In the second group ($G2$), all objects detected twice are selected. Then, in the last group ($G3$), only objects detected once are selected. After detection, the age is updated, where we assign each group a different a as specified in Equation 3.

For implementation, we used the histogram of oriented gradients (HOG) together with an SVM as proposed by [15]. More specifically, we used the DLIB implementation from [31]. The HOG features rely on cells which make them either inaccurate (on pixel level) or consume large amounts of memory (overlapping cells). In our implementation, we shifted the computed gradients below the cell grid in x and y directions ranging from one to eight pixels (used cell size cs). For each shift, we run a detection and collect the results. The idea is that the average of all detections is accurate. For some of those detections, the counterpart is missing (no detection on the opposite shift); therefore, we perform outliers removal for two times the standard deviation. The shift procedure not only improves the accuracy, but also increases the detection rate.

Another issue with accuracy is when it comes to deformable objects in addition to moving occlusions, changing lighting conditions, and distractors (Figure 3). Specifically, for pupil center detection tasks, the circular pupil deforms to an ellipse as shown in Figure 3. Moreover, the pupil size changes and many



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Subset of challenges which arise in pupil center detection. Deformations, reflections, motion blur, nearly closed eyes, and contact lenses are shown. Images are taken from [21, 18, 20].

people use makeup or need eyeglasses: All of which lead to reflections in the near infrared spectrum. To adapt to those challenges, we propose to use a grid of detectors and average over the deformation. This averaging is dependent on the combination possibilities for different types of success patterns of the grid (symmetric patterns).

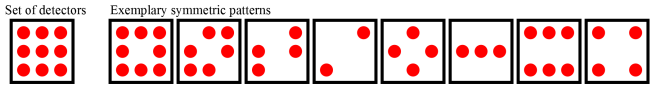


Fig. 4. Some exemplary symmetric means for a detector grid with size nine.

In our implementation, we chose the minimal grid consisting of nine detectors with a shift of gs pixels. Some valid symmetric mean patterns can are shown in Figure 4, where a red dot indicates that the detector belonging to this grid position found an object. Those patterns can be calculated using the binomial coefficient to get all possible combinations. For evaluation, if it is symmetric, the sum of coordinates has to be zero if they are centered on the central detector (for example $x, y \in \{-1, 0, 1\}$ where $(0, 0)$ is the central detector).

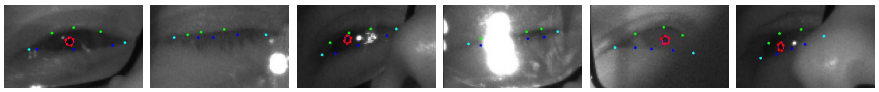
4 New Dataset

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 5. Exemplary images of the new dataset.

In addition to the proposed algorithm, this work contributes a new dataset with more than 16,200 hand-labeled images (1280×752) from six different subjects. These images were recorded using a near-infrared remote camera in a

driving simulator setting (prototype car with all standard utilities included) at Bosch GmbH, Germany. As exemplary shown in Figure 5, the subjects drove in a naturalistic way, e.g., when turning the steering wheel, eyes or head are occluded.



(a) (b) (c) (d) (e) (f)

Fig. 6. Exemplary eyelid and pupil annotations. The red dots are on the pupil boundary, green dots represent the upper eyelid, blue dots the lower eyelid, and the turquoise dots are on the eye corners.

We annotated all eyes on these images using a modified version of EyeLad from [19]. Eyes that are occluded by approximately 50% were not annotated. We labeled the smallest enclosing eye boxes: The pupil outline with five points, and for the eye corners and the upper and lower eyelid, we used three points each. The pupil annotation consists of five points on the outline with sub-pixel accuracy (Figure 6). This new data contains different kind of occlusions: For instance, reflections (Figure 6(d)), the nose of the subject (Figure 6(f)), occlusion due to steering (Figure 6(e)), and occlusion of the pupil or eyelids due to eyelashes (Figure 6(b)). Therefore, we believe that our data set is a valuable contribution to the research community in the realm of object detection, specifically for gaze tracking.

Table 1. Eye detection results (recall; T=true, F=false) for the first, middle and last iteration. Subject 6 (images on the left) has many unannotated frames, since eyes are occluded by approximately 50% (100% of the error is on non-annotated locations). The red star represents a detection by our algorithm that was not annotated and the green star represents an annotation that was successfully found.

		Detector						10 annotations					
Dataset	Subject	First		Mid		Last		First		Mid		Last	
		T	F	T	F	T	F	T	F	T	F	T	F
Proposed	Sub1	.99	0	1	0	1	0	.95	0	1	0	1	0
	Sub2	.94	0	1	.01	1	.01	.59	0	.90	.01	1	.01
	Sub3	.71	.01	.96	.02	.97	.02	.30	0	.85	.06	.95	.02
	Sub4	.99	0	.99	0	.99	0	.78	0	.99	0	.99	0
	Sub5	.60	0	.93	.03	.98	.02	.46	.01	.82	.03	.97	.03
	Sub6	.59	.01	.91	.03	.98	.09	.73	.01	.99	.14	1	.28
GI4E [55]		.36	0	.95	0	.96	0	.22	0	.58	0	.92	0
[17]		.43	0	.55	.01	.92	.03	.48	0	.84	.02	.93	.04

Table 2. Head mounted pupil center detection results error up to five pixels [21].

Dataset	ExCuSe [18]																		ElSe [21]				
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Swirski	.05	.23	.06	.34	.78	.19	.39	.41	.23	.30	.20	.71	.61	.51	.62	.18	.66	.15	.09	.22	.08	.02	.96	.43
ExCuSe	.71	.39	.37	.79	.75	.59	.48	.55	.75	.78	.58	.79	.69	.68	.55	.34	.78	.23	.23	.57	.52	.26	.93	.45
ElSe	.85	.65	.63	.83	.84	.77	.59	.68	.86	.78	.75	.79	.73	.84	.57	.59	.89	.56	.33	.78	.47	.52	.94	.52
Proposed	**.89**	**.81**	**.79**	**.93**	**.93**	**.89**	**.82**	**.88**	**.90**	**.93**	**.94**	**.88**	**.84**	**.91**	**.69**	**.92**	**.98**	**.62**	**.53**	**.89**	**.82**	**.73**	**.98**	**.69**

Table 3. Remote pupil center detection results (3 and 6 are the pixel error).

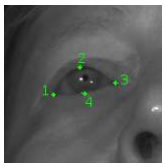
		GI4E [55]	BioID [27]	[17]	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6
ElSe	3	.07	.16	.26	0	.45	.01	0	.01	0
	6	.50	.43	.63	.04	.67	.14	.06	.14	0
Proposed	3	**.94**	**.85**	**.64**	**.93**	**.83**	**.82**	**.95**	**.92**	**.61**
	6	**.98**	**.93**	**.81**	**.98**	**.99**	**.90**	**.96**	**.98**	**.71**

5 Evaluation

We evaluated our algorithm on several publicly available data sets ([55, 17, 64, 18, 21, 17, 27]) for self learning together with our proposed dataset. The first evaluation is without the grid of detectors to demonstrate the performance of the aging approach itself. Table 1 shows the results for the eye detection task (without grid). We ran the algorithm for a maximum of 15 iterations. Most of the error in the proposed data set stems from unlabeled images due to the annotation criteria of labeling only eyes with less than 50% occlusion. This error is apparent especially for subject 6, where the error reaches 28% in relation to all possible correct detections. The same applies for subject 2 and 5. The subsequent evaluations refer to pixel-precise object recognition.

Table 2 shows the results for comparing our approach to the state-of-the-art algorithms [49], ExCuSe [18], and ElSe [21]. The results support that our approach, for all datasets, had the highest detection performance. Here, the maximum of iterations was set to 15. For initialization of our algorithm, we selected ten annotations. The distance between the selected annotations was again ten frames ($i \mod 10 = 0$). Though our algorithm outperforms all the competitors, the results provide a basis for even further improvement. The input to the algorithm was each entire data set, except for data set XIX. Here, we performed the same selection of ten frames from 13,473 images as with the other sets, but for the iterations, we divided it into three sets. They were set sizes of

Table 4. Remote eyelid point detection results (3 and 6 are the pixel error).



	Proposed								[30]							
	Left		Right		Upper		Lower		Left		Right		Upper		Lower	
	3	6	3	6	3	6	3	6	3	6	3	6	3	6	3	6
Sub1	**.91**	.98	**.87**	**.97**	**.31**	**.50**	**.80**	**.99**	.88	**.99**	.01	.21	.29	.48	.32	.94
Sub2	.75	.95	**.66**	**.89**	**.43**	**.70**	**.69**	**.96**	**.77**	**.99**	.28	.82	.10	.27	.56	**.96**
Sub3	**.64**	**.90**	**.59**	**.88**	**.35**	**.61**	**.68**	.93	.28	.77	.32	.57	.18	.45	.64	**.94**
Sub4	**.46**	**.93**	**.82**	**.98**	**.34**	**.72**	**.80**	**.98**	.39	.76	.39	.73	.03	.12	.66	.94
Sub5	**.46**	**.73**	**.58**	**.91**	**.30**	**.61**	**.63**	.79	.43	**.73**	.41	.65	.18	.45	.54	**.85**
Sub6	.29	.58	**.31**	**.53**	**.30**	**.60**	.40	.68	**.34**	**.69**	.24	.52	.23	.54	**.52**	**.83**

5,000 and 3,473 images for the first two sets and the last set respectively. This division was made due to the original size of the data set exceeding the memory capacity of our server.[3]

For comparison in remote pupil detection, we chose the best competitor in [17], which is the second part of ElSe [21], since it outperformed all the other algorithms [9, 53, 22] on all datasets. For data sets GI4E [55], BioID [27], and [17], we used the labeled eye boxes and increased the size by twenty pixels in each direction: In order to increase the challenge. For the proposed dataset, we selected the eye center and extracted a 161×161 area surrounding it. We only used the left eye (from the viewer perspective) for the pupil center evaluation to reduce the data set size. For the proposed approach, we initially selected again ten images with a fixed distance of ten ($i \mod 10 = 0$). As indicated in Table 3, the proposed approach surpasses the state-of-the-art. Moreover, the effect of the increased eye boxes is shown for ElSe.[4]

For the eyelid experiment, we evaluated our approach against the shape detector from [30]. This predictor was trained on all data sets except the one for evaluation; for example, the evaluation for subject 1 involved training the predictor on subjects 2 through 6. The defined eyelid shape is constructed by four points as illustrated in the image next to Table 4. The left and right eye corner points are used as the ground truth data. For the upper and lower eyelid point, we interpolated the connection using Bezier splines and selected the center point on both curves. The images were the same as in the previous experiment. For the point selection, we again used ten points with distance ten ($i \mod 10 = 0$). We selected different starting locations to give a more broad spectrum of possible results of the algorithm. As can be seen in Table 4, our algorithm is more often the most accurate, even for the condition to detect each point separately without any global optimization between the points. In addition, it should be noted that we optimize the evaluation for the approach from [30]. This means that [30] expects to receive an equally centered bounding box on the object to estimate the outline, otherwise it fails. For our approach, it does not change anything if the eye box is shifted.[4]

6 Conclusion

We proposed a novel algorithm for automatic and accurate point labeling in various scenarios with remarkable performance. While our algorithm is capable of generating detectors in addition to the annotation, it remains difficult to evaluate their generality: Hence, we refer to them as specialized detectors. In addition to the proposed algorithm, we introduced a dataset with more than 16,000 manually labeled images with annotated eye boxes, eye lid points, eye corner, and the pupil outline, which will be made publicly available together with a library.

[3] Parameters: detection window size 65×65, cell size $cs = 8$, the grid shift $gs = 5$; SVM: $\epsilon = 0.01$, $C = 1$

[4] Parameters: detection window size 31×31, cell size $cs = 4$, the grid shift $gs = 2$; SVM: $\epsilon = 0.01$, $C = 1$

References

1. Arora, N., Allenby, G.M., Ginter, J.L.: A hierarchical bayes model of primary and secondary demand. Marketing Science **17**(1), 29–44 (1998)
2. Bakker, B., Heskes, T.: Task clustering and gating for bayesian multitask learning. Journal of Machine Learning Research **4**(May), 83–99 (2003)
3. Baxter, J.: A model of inductive bias learning. J. Artif. Int. Res. **12**(1), 149–198 (Mar 2000), http://dl.acm.org/citation.cfm?id=1622248.1622254
4. Bertozzi, M., Broggi, A.: Gold: A parallel real-time stereo vision system for generic obstacle and lane detection. IEEE transactions on image processing **7**(1), 62–81 (1998)
5. Braunagel, C., Rosenstiel, W., Kasneci, E.: Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. IEEE Intelligent Transportation Systems Magazine (2017)
6. Cao, X., Wang, Z., Yan, P., Li, X.: Transfer learning for pedestrian detection. Neurocomputing **100**, 51–57 (2013)
7. Caruana, R.: Multitask learning. In: Learning to learn, pp. 95–133. Springer (1998)
8. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence **34**(4), 743–761 (2012)
9. Droege, D., Paulus, D.: Pupil center detection in low resolution images. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. pp. 169–172. ACM (2010)
10. Duchowski, A.T.: Eye tracking methodology. Theory and practice **328** (2007)
11. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)
12. Farajidavar, N., de Campos, T.E., Kittler, J.: Adaptive transductive transfer machine. In: BMVC (2014)
13. FarajiDavar, N., De Campos, T., Kittler, J., Yan, F.: Transductive transfer learning for action recognition in tennis games. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 1548–1553. IEEE (2011)
14. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence **28**(4), 594–611 (2006)
15. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2010)
16. Fuhl, W., Santini, T., Reichert, C., Claus, D., Herkommer, A., Bahmani, H., Rifai, K., Wahl, S., Kasneci, E.: Non-intrusive practitioner pupil detection for unmodified microscope oculars. Computers in Biology and Medicine **79**, 36–44 (12 2016)
17. Fuhl, W., Geisler, D., Santini, T., Rosenstiel, W., Kasneci, E.: Evaluation of state-of-the-art pupil detection algorithms on remote eye images. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. pp. 1716–1725. ACM (2016)
18. Fuhl, W., Kübler, T., Sippel, K., Rosenstiel, W., Kasneci, E.: ExCuSe: Robust Pupil Detection in Real-World Scenarios, pp. 39–51. Springer International Publishing, Cham (2015)
19. Fuhl, W., Santini, T., Geisler, D., Kübler, T., Kasneci, E.: Eyelad: Remote eye tracking image labeling tool (02 2017)

20. Fuhl, W., Santini, T., Kasneci, G., Kasneci, E.: Pupilnet: Convolutional neural networks for robust pupil detection. CoRR **abs/1601.04902** (2016)
21. Fuhl, W., Santini, T.C., Kuebler, T., Kasneci, E.: Else: Ellipse selection for robust pupil detection in real-world environments. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. pp. 123–130. ETRA '16, ACM, New York, NY, USA (2016)
22. George, A., Routray, A.: Fast and accurate algorithm for eye localization for gaze tracking in low resolution images. arXiv preprint arXiv:1605.05272 (2016)
23. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: European conference on computer vision. pp. 392–407. Springer (2014)
24. Hoffman, J., Guadarrama, S., Tzeng, E.S., Hu, R., Donahue, J., Girshick, R., Darrell, T., Saenko, K.: Lsda: Large scale detection through adaptation. In: Advances in Neural Information Processing Systems. pp. 3536–3544 (2014)
25. Huang, J., Gretton, A., Borgwardt, K.M., Schölkopf, B., Smola, A.J.: Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems. pp. 601–608 (2007)
26. Jain, V., Learned-Miller, E.: Online domain adaptation of a pre-trained cascade of classifiers. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 577–584. IEEE (2011)
27. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: International Conference on Audio-and Video-Based Biometric Person Authentication. pp. 90–95. Springer (2001)
28. Kasneci, E., Hardiess, G.: Driving with Homonymous Visual Field Defect. In Homonymous Visual Field Defects. Springer International Publishing (2017)
29. Kasneci, E., Kuebler, T., Broelemann, K., Kasneci, G.: Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth. Computers in Human Behavior, Elsevier **68**, 450–455 (03 2017)
30. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (2014)
31. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**(Jul), 1755–1758 (2009)
32. Kübler, T.C., Rittig, T., Kasneci, E., Ungewiss, J., Krauss, C.: Rendering refraction and reflection of eyeglasses for synthetic eye tracker images. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. pp. 143–146. ETRA '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2857491.2857494, http://doi.acm.org/10.1145/2857491.2857494
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
34. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 2200–2207 (2013)
35. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. IEEE transactions on pattern analysis and machine intelligence **28**(11), 1863–1868 (2006)
36. Nakajima, C., Pontil, M., Heisele, B., Poggio, T.: Full-body person recognition system. Pattern recognition **36**(9), 1997–2006 (2003)

37. Namin, S.T., Najafi, M., Salzmann, M., Petersson, L.: A multi-modal graphical model for scene analysis. In: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. pp. 1006–1013. IEEE (2015)
38. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583), 607–607 (1996)
39. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks **22**(2), 199–210 (2011)
40. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
41. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: International Conference on Machine Learning. pp. 82–90 (2014)
42. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on Machine learning. pp. 759–766. ACM (2007)
43. Raj, A., Namboodiri, V.P., Tuytelaars, T.: Subspace alignment based domain adaptation for rcnn detector. arXiv preprint arXiv:1507.05578 (2015)
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
45. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
46. Sugiyama, M., Krauledat, M., MÃžller, K.R.: Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research **8**(May), 985–1005 (2007)
47. Sun, Q., Chattopadhyay, R., Panchanathan, S., Ye, J.: A two-stage weighting framework for multi-source domain adaptation. In: Advances in neural information processing systems. pp. 505–513 (2011)
48. Świrski, L., Dodgson, N.: Rendering synthetic ground truth images for eye tracker evaluation. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 219–222. ETRA '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2578153.2578188, http://doi.acm.org/10.1145/2578153.2578188
49. Świrski, L., Bulling, A., Dodgson, N.: Robust real-time pupil tracking in highly off-axis images. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 173–176. ACM (2012)
50. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
51. Tafaj, E., Kasneci, G., Rosenstiel, W., Bogdan, M.: Bayesian online clustering of eye movement data. In: Proceedings of the Symposium on Eye Tracking Research and Applications. pp. 285–288. ACM (2012)
52. Thrun, S., Pratt, L.: Learning to learn. Springer Science & Business Media (2012)
53. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. VISAPP **11**, 125–130 (2011)
54. Ullman, S.: High-level vision: Object recognition and visual cognition, vol. 2. MIT press Cambridge, MA (1996)
55. Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., Cabeza, R.: Hybrid method based on topography for robust detection of iris center and eye corners.

ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **9**(4), 25–25 (2013)
56. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. vol. 1, pp. I–I. IEEE (2001)
57. Wan, J., Ruan, Q., Li, W., Deng, S.: One-shot learning gesture recognition from rgb-d data using bag of features. The Journal of Machine Learning Research **14**(1), 2549–2582 (2013)
58. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
59. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. pp. 131–138. ETRA '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2857491.2857492, http://doi.acm.org/10.1145/2857491.2857492
60. Wu, D., Zhu, F., Shao, L.: One shot learning gesture recognition from rgbd images. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. pp. 7–12. IEEE (2012)
61. Xu, J., Ramos, S., Vázquez, D., López, A.M., Ponsa, D.: Incremental domain adaptation of deformable part-based models. In: BMVC (2014)
62. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. In: Advances in Neural Information Processing Systems. pp. 1790–1798 (2014)
63. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. CoRR **abs/1611.08860** (2016), http://arxiv.org/abs/1611.08860
64. Zhou, F., Brandt, J., Lin, Z.: Exemplar-based graph matching for robust facial landmark localization. In: IEEE International Conference on Computer Vision (ICCV) (2013)