

Hand Pose Estimation Using Hierarchical Detection

B. Stenger^{1*}, A. Thayananthan¹, P.H.S. Torr², and R. Cipolla¹

¹ University of Cambridge, Department of Engineering
Cambridge, CB2 1PZ, UK
{bdrs2,at315,cipolla}@eng.cam.ac.uk

² Oxford Brookes University, Department of Computing
Oxford OX33 1HX, UK
philiptorr@brookes.ac.uk

Abstract. This paper presents an analysis of the design of classifiers for use in a hierarchical object recognition approach. In this approach, a cascade of classifiers is arranged in a tree in order to recognize multiple object classes. We are interested in the problem of recognizing multiple patterns as it is closely related to the problem of locating an articulated object. Each different pattern class corresponds to the hand in a different pose, or set of poses. For this problem obtaining labelled training data of the hand in a given pose can be problematic. Given a parametric 3D model, generating training data in the form of example images is cheap, and we demonstrate that it can be used to design classifiers almost as good as those trained using non-synthetic data. We compare a variety of different template-based classifiers and discuss their merits.

1 Introduction

This paper considers the problem of locating and tracking an articulated object using a single camera. The method is illustrated by the problem of hand detection and tracking. There is a lot of ambiguity in the problem of tracking a complex articulated object, and a successful method should be able to maintain multi-modal distributions over time. A number of different techniques have been suggested to deal with multi-modality, e.g. particle filtering [7, 11]. When track is lost, a robust tracker should devise a recovery strategy, as for example in region based tracking [23]. This task, however, can be seen as a detection problem, and thus in [19, 20] it is argued that the tracking of complex objects should involve the close synthesis of object detection and tracking.

Object recognition is typically considered as the task of detecting a single class of objects $\mathcal{O}$ (e.g. faces) in a scene $\mathcal{I}$; locating the object in the scene and determining its pose. But suppose we are interested in recognizing m categories of objects $\mathcal{O}_1, \dots, \mathcal{O}_m$ simultaneously, i.e. are any of a set of objects in the scene, and if so where? How can it efficiently be decided whether the scene contains

* The author is currently with Toshiba Research, Kawasaki, Japan.

one of these objects? One option that is commonly followed is to independently train a classifier for each object [16]. The drawback of such an approach is that computation time scales roughly linearly in the number of objects to be identified. Recently advances have been made in face detection based on the idea of a cascade of classifiers [15, 22], where successively more complex classifiers are combined in a cascade structure, which increases the speed of the detector by focusing attention on promising regions of the image, see figure 1(a). First the image is divided into a set of subregions. The initial classifier eliminates a large portion of these subwindows with little computation; those remaining are processed further down the cascade. At each level the number of subwindows remaining decreases, allowing for more computationally expensive classifiers to be used at the bottom level for accurate discrimination of the remaining subwindows. As the motivation for such a cascade is the minimization of computation time, this paper examines some of the issues involved in classifier design for efficient template-based classification for hand pose estimation.

The next section reviews related work on hierarchical detection and describes the links to 3D pose estimation. Section 3 introduces the shape and colour features that are used, as well as the templates for classification. An evaluation of these classifiers in terms of performance and efficiency is presented in section 4, and their application to a pose estimation problem is shown.

2 Pose Estimation Using Shape Templates

One question is whether a cascaded approach can be used for recognizing multiple objects, i.e. how to design a computationally efficient cascade of classifiers for a given set of objects, $\mathcal{O}_1, \dots, \mathcal{O}_m$. The problem is closely linked to the recognition of articulated objects which can be thought of as an infinite collection of objects indexed by the joint articulation parameters. In particular, Gavrilu [6] examines the problem of detecting pedestrians. Chamfer matching [2] is used to detect humans in different poses, and detecting people is formulated as a template matching problem. When matching many similar templates to an image, a significant speed-up can be achieved by forming a template hierarchy and using a coarse to fine search [6, 14]. Toyama and Blake [21] use exemplar templates to evaluate likelihoods within a probabilistic tracking framework. Shape templates of a walking person are clustered and only the chamfer cost of the prototypes needs to be computed. However, with increasing object complexity the number of exemplars required for tracking rises as well. If a parametric 3D object model is available, the generation of training examples is cheap. Additionally, each generated 2D template is annotated with the 3D model parameters, thus pose recovery can be formulated as object detection: create a database of model-generated images and use a nearest-neighbour search to find the best match. This approach is followed, for example, by Athitsos and Sclaroff for hand pose estimation [1] and Shakhnarovich *et al.* [17] for upper body pose estimation. In [19] it is suggested to partition the parameter space of a 3D hand model using a multi-resolution grid. A distribution is defined on the finest grid and is propagated over time.

Algorithm 1 : Cascade of Classifiers for Multiple Categories

```

for each subwindow I
  start at root node  $(l, k) = (1, 1)$ .
  if  $C_k^l(\mathbf{I}) > 0$  then repeat for child nodes of  $(l, k)$ .
  else assign zero probability to all child nodes of  $(l, k)$ .
  endif
endfor

```

This has the advantage that temporal information can be used to resolve ambiguous situations and to smooth the motion. Shape templates, generated by the 3D model, are used to evaluate the likelihoods in regions of the state space. The templates are arranged in a hierarchy and are used to rapidly discard regions with low probability mass. For the first frame, the tree corresponds to a detection tree, thus the idea of cascaded classifiers can be applied, which eliminate large regions of the parameter space at early stages and focus computation on ambiguous regions. In terms of classifiers, the aim is to maintain a high detection rate while rejecting as many false positives as possible at each node in the tree. Within this paper we will analyze the design of cascaded classifiers for such a hierarchy, which can be used within the tree-based filtering framework of [19].

Given a tree which at each level partitions the set of models into mutually exclusive regions $\mathcal{S}_i^l$ for $l = 1, \dots, L$ where L is the number of levels in the tree and $i = 1 \dots N_L$ where N_L is the number of sets at that level. So that $\mathcal{S} = \cup_i \mathcal{S}_i^l$ and $\mathcal{S}_j^l \cap \mathcal{S}_k^l = 0; \forall j, k$. The goal is to design a classifier $\mathcal{C}_j^l$ which achieves high detection rates with modest false positive rates on the region $\mathcal{S}_j^l$. The search then proceeds as shown in algorithm 1. A schematic of this algorithm is shown in figure 1b. The next section examines a number of different classifiers based on edge and colour features.

3 Explanation of Features and Classifiers

Edges (occluding contours) and colour (silhouette interior) have proved useful features for recognizing hands and discriminating between different poses, e.g. [1, 13]. Each feature is treated independently, assuming that in different settings one of the features may be sufficient for recognition. Edge features are considered in the following section.

3.1 Edge Features

When using edges as features, robust similarity functions need to be used when comparing a template with the image, i.e. ones that are tolerant to small shape changes. One way to achieve this is to blur the edge image or template before correlating them. Other methods, which are tolerant to small shape deformations and some occlusion are the (truncated) chamfer and Hausdorff distance functions [2, 10]. Both methods are made efficient by the use of fast operations like

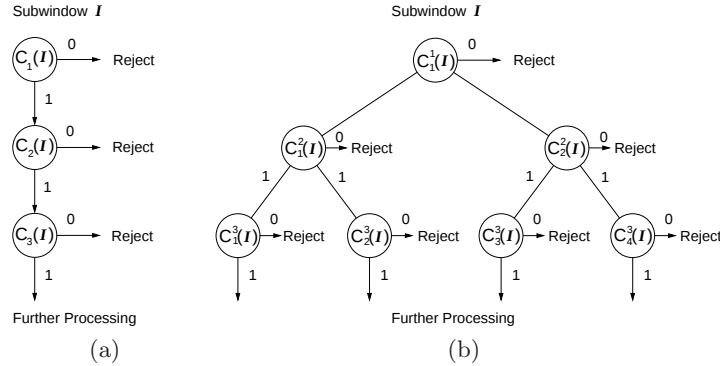


Fig. 1. Cascade of Classifiers (a) A cascade of classifiers for a single object class where each classifier has a high detection and moderate false positive rate (b) Classifiers in a tree structure; in a tree-based object recognition scheme each leaf corresponds to a single object class. When objects in the subtrees have similar appearance, classifiers can be used to quickly prune the search. A binary tree is shown here, but the branching factor can be larger than two

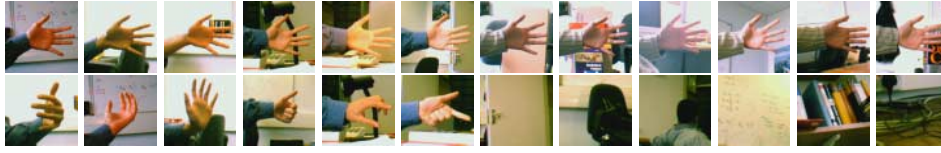


Fig. 2. Examples of Training and Test Images. **Top:** First six images from left: training images, last six: test images **Bottom:** First six images from left: negative training examples containing hand in different pose, last six: negative examples containing background

the distance transform or dilation of the edge image. Olson and Huttenlocher [14] include edge orientation in Hausdorff matching. This is done by decomposing both template and edge image into a number of separate channels according to edge orientation. The distance is computed separately for each channel, and the sum of these is the total cost.

Both chamfer and Hausdorff matching can be viewed as special cases of linear classifiers [5]. Let $\mathbf{B}$ be the feature map of an image region, for example a binary edge image. The template $\mathbf{A}$ is of the same size and can be thought of as a prototype shape. The problem of recognition is to decide whether or not $\mathbf{B}$ is an instance of $\mathbf{A}$. By writing the entries of matrices $\mathbf{A}$ and $\mathbf{B}$ into vectors $\mathbf{a}$ and $\mathbf{b}$, respectively, this problem can be written as a linear classification problem with a discriminant function $\mathbf{a}^T \mathbf{b} = c$, with constant c . This generalization also permits negative coefficients of $\mathbf{a}$, potentially increasing the cost of cluttered image areas, and different weights may be given to different parts of the shape. Felzenszwalb [5] has shown that a single template $\mathbf{A}$ and a dilated edge map $\mathbf{B}$ is sufficient to detect a variety of shapes of a walking person. A classifier is thus

defined by the entries in the matrix $\mathbf{A}$ and in this paper the following classifiers are evaluated (illustrated in figure 3 (a)). For each type two sets of templates are generated, one with and one without orientation information. For oriented edges the angle space is subdivided into six discrete intervals, resulting in a template for each orientation channel.

Centre Template: This classifier uses a single shape template $\mathbf{A}$, generated using the centre of a region in parameter space. Two possibilities for the feature matrix $\mathbf{B}$ are compared. One is the distance transformed edge image in order to compute the truncated chamfer distance [6]. For comparison, the Hausdorff fraction is computed using the dilated edge image [9]. The parameters for both methods are set by testing the classification performance on a test set of 5000 images. Values for the chamfer threshold τ from 2 to 120 were tested, and $\tau = 50$ was chosen, but little variation was observed for values larger than 20. For the dilation parameter δ values from 1 to 11 were compared, and $\delta = 3$ showed the best performance.

Marginalized Template: In order to construct a classifier which is sensitive to a particular region in parameter space, the template $\mathbf{A}$ is constructed by densely sampling the values in this region, and simultaneously setting the model parameters to these values. The resulting model projections are then pixel-wise added and smoothed. Different versions of matrices $\mathbf{A}$ are compared: (a) the pixel-wise average of model projections, (b) the pixel-wise average, additionally setting the background weights uniformly to a negative value such that the sum of coefficients is zero, and (c) the union of all projections, resulting in a binary template.

Linear Classifier Learnt from Image Data: The template $\mathbf{A}$ is obtained by learning a classifier as described by Felzenszwalb [5]. A labelled training set containing 1,000 positive examples and 4,000 negative examples of which 1,000 contain the hand in a different pose and 3,000 images contain background regions (see figure 2) is used to train a linear classifier by minimizing the perceptron cost function [4].

3.2 Colour Features

Given an input image region, define the feature matrix $\mathbf{B}^s$ as the log-likelihood map of skin colour and $\mathbf{B}^{bg}$ as the log-likelihood map of background colour. Skin colour is represented as a Gaussian in (r, g) -space, and the background distribution is modelled as a uniform distribution. A silhouette template $\mathbf{A}$ is defined as containing +1 at locations within the hand silhouette and zero otherwise. Writing these matrices as vectors, a cost function, which corresponds to the log-likelihood [18] can be written as:

$$\mathbf{a}^T(\mathbf{b}^s - \mathbf{b}^{bg}) + \mathbf{1}^T \mathbf{b}^{bg} \quad , \quad (1)$$

where the entries in $\mathbf{b}^{bg}$ are constant when using a uniform background model. Thus, correlating the matrix $\mathbf{B} = \mathbf{B}^s - \mathbf{B}^{bg}$ with a silhouette template $\mathbf{A}$ corresponds to evaluating the log-likelihood of an input region up to an additive

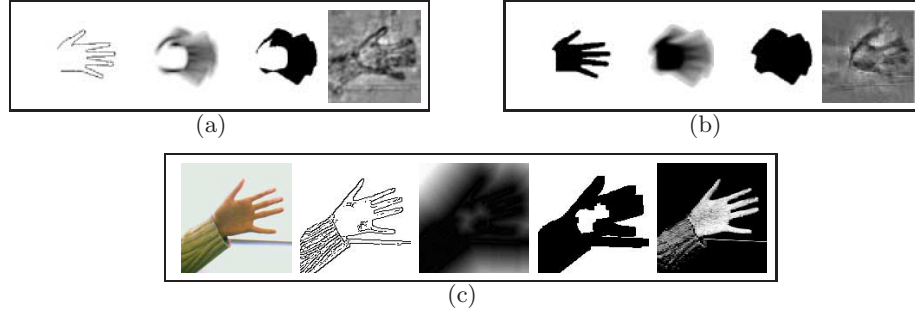


Fig. 3. Templates and Feature Maps used for Classification (a) Templates used for classifying edge features. From left to right: single template, marginalized template, binary marginalized template, template learnt from image data (b) Templates for classifying colour features, analogous to top row (c) Extracted features from an image with simple background. From left to right: input image, edges, distance transformed edge map, dilated edge map, colour likelihood image

constant. Note that when the distributions are fixed, the log-likelihood values for each colour vector can be pre-computed and stored in a look-up table beforehand. The corresponding templates **A** are shown in figure 3 (b).

4 Results

In order to compare the performance of different classifiers, a labelled set of hand images was collected. Positive examples are defined as the hand being within a region in parameter space. For the following experiments this region is chosen to be a rotation of 30 degrees parallel to the image plane. Negative examples are background images as well as images of hands in configurations outside of this parameter region. The evaluation of classifiers is done in three independent experiments for different hand poses, an open hand, a pointing hand and a closed hand. The test data sets each contain 5000 images, of which 1000 are true positive examples. The classifiers are defined by the entries in the matrix **A**, described in the previous section, and illustrated in figure 3 (a) and (b).

4.1 Edge Templates

The following observations were made consistently in the experiments:

- In all cases the use of edge orientation resulted in better classification performance. Including the gradient direction is particularly useful when discriminating between positive examples and negative examples of hand images. This is illustrated in figure 4 (a) and (b), which show the class distributions for non-oriented edges and oriented edges in the case of marginalized templates with non-negative weights. The corresponding ROC curves are shown

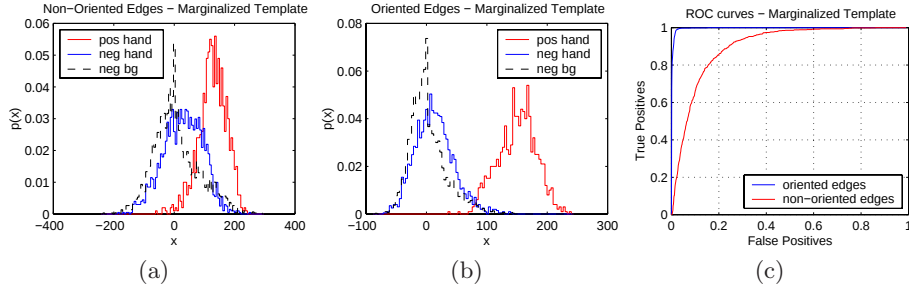


Fig. 4. Including Edge Orientation Improves Classification Performance

This example shows the classification results on a test set using a marginalized template (a) histogram of classifier output using edges without orientation information: hand in correct pose (red/light), hand in incorrect pose (blue) and background regions (black,dashed line) (b) histogram of classifier output using edges with orientation information. The classes are clearly better separated, (c) the corresponding ROC curve

in 4 (c), demonstrating the benefit of using oriented edges. This shift in the ROC curve is observed in different amounts for all classifiers and is shown in figure 5 (a) and (b).

- In all experiments the best classification results were obtained by the classifier trained on real data. The ROC curves for a particular hand pose (open hand parallel to image plane) are shown in figure 5. At a detection rate of 0.99 the false positive rate was below 0.05 in all experiments.
- Marginalized templates showed good results, also yielding low false positive rates at high detection rates. Templates using pixel-wise averaging and negative weights for background edges were found to perform best when comparing the three versions of marginalized templates. For this template the false positive rates were below 0.11 at a detection rate of 0.99.
- Using the centre template with chamfer or Hausdorff matching showed slightly lower classification performance than the other methods, but in all cases the false positive rate was still below 0.21 for detection rates of 0.99. Chamfer matching gave better results than Hausdorff matching, as can be seen in the ROC curve in figure 5 (b). It should be noted that this result is in contrast to the observations made by Huttenlocher in [8], for more details see [18].

The execution times for different choices of templates $\mathbf{A}$ were compared in order to assess the computational efficiency. Computing the scalar product of two vectors of size 128×128 is relatively expensive. However, the computational time can be reduced by avoiding the multiplication of zero valued entries in the matrix $\mathbf{A}$. For chamfer and Hausdorff matching, the template only contains the points of a single model projection. The number of points in the marginalized template depends on the size of the parameter space region it represents. In the experiments it contained approximately 14 times as many non-zero points as a single model template. When using a binary template the dot product

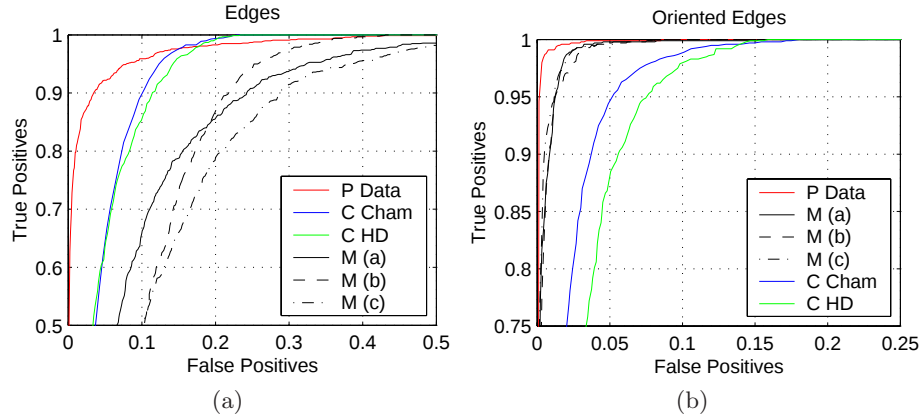


Fig. 5. ROC Curves For Classifiers This figure shows the ROC curve for each of the classifiers (a) edge features alone, and (b) oriented edges. Note the difference in scale of the axes. The classifier trained on real image data performs best, the marginalized templates all show similar results, and chamfer matching is slightly better than Hausdorff matching in this experiment. When used within a cascade structure, the performance at high detection rates is important.

computation simplifies to additions of coefficients. If both vectors are in binary form, a further speed-up can be achieved by using AND operations [18]. The execution times for correlating 10,000 templates are shown in table 1. The time for computing a distance transform or dilation, which needs to be only computed once for each frame when chamfer or Hausdorff matching is used, is less than 2 ms and is therefore negligible when matching a large number of templates. There clearly seems to be a trade-off between computation time and classification performance for the classifiers. When used in a cascaded structure, the detection rate of a classifier needs to be very high, so as not to miss any true positives. Chamfer and Hausdorff matching, while having a larger false positive rate, are about 10-14 times faster to evaluate than marginalized templates and about 40 times faster than the trained classifier.

4.2 Silhouette Templates

The same test data set as for edges was used, and the following observations were made:

- For the test set colour information helps to discriminate between positive examples of hands and background regions. However, there is significant overlap between the positive and negative class examples which contain a hand. Oriented edges are better features to discriminate between the hand in different poses, whereas colour features are slightly better at discriminating between the positive class and background regions.

Table 1. Computation times for Correlating Templates The execution times for computing the dot product of 10,000 image patches of size 128×128 , where only the non-zero coefficients are correlated for efficiency, measured on a 2.4 GHz PC with Pentium IV. The last column shows the false positive rates for each classifier at a fixed detection rate of 0.99

Classification Method	Number of Points	Execution Time	fp at $tp = 0.99$
Chamfer	400	13 ms	0.10
Hausdorff	400	13 ms	0.12
Marginalized Template	5,800	186 ms	0.02
Binary Marginalized Template	5,800	136 ms	0.02
Trained Classifier Template	16,384	524 ms	0.01

- Both, centre template and marginalized template show better classification performance than the trained classifier, in particular in the high detection range. At detection rates of 0.99 the false positive rate for the centre template is 0.24, whereas it is 0.64 for the trained classifier. However, the trained classifier shows better performance at separating positive examples from negative example images containing hands. At a detection rate of 0.99, the false positive rate is 0.41 compared to 0.56 for the other two classifiers.

The evaluation can be performed efficiently by pre-computing a sum table, $\mathbf{B}^{sum}$, which contains the cumulative sums of costs along the x -direction:

$$\mathbf{B}^{sum}(x, y) = \sum_{i=1}^x (\log p^s(I(i, y)) - \log p^{bg}(I(i, y))) \quad , \quad (2)$$

where in this equation the image I is indexed by its x and y -coordinates. p^s and p^{bg} are the skin colour and background colour distributions, respectively. This array only needs to be computed once, and is then used to compute sums over areas by adding and subtracting values at points on the silhouette contour. It is a convenient way to convert area integrals into contour integrals, and is related to the summed area tables of Crow [3], the integral image of Viola and Jones [22] or the integration method based on Green's theorem of Jermyn and Ishikawa [12]. Compared to integral images the sum over non-rectangular regions can be computed more efficiently, and in contrast to the technique of Jermyn and Ishikawa the contour normals are not needed. In contrast to these methods, however, the model points need to be connected pixels, e.g. obtained by line-scanning a silhouette image. The computation time for evaluating 10,000 templates is reduced from 524 ms to 13 ms for a silhouette template of 400 points. As the computation of the sum table $\mathbf{B}^{sum}$ is only computed once, this is negligible when matching a large number of templates.



Fig. 6. Hierarchical Detection of Pointing Hand Left: Input image, Next: Images with classification results super-imposed. Each square represents an image location which contains at least one positive classification result. Higher intensity indicates larger number of matches. Face and second hand introduce ambiguity. Regions are progressively eliminated, the best match is shown on the right

	Accepted				Rejected			
Level 1								
	9.84	11.08	12.21	12.29	22.90	22.90	22.90	22.90
Level 2								
	7.36	8.41	9.63	9.69	17.00	17.01	17.02	17.07
Level 3								
	6.68	6.69	6.85	6.90	12.21	12.24	12.25	13.21

Fig. 7. Search Results at Different Levels of the Tree This figure shows typical examples of accepted and rejected templates at levels 1 to 3 of the tree, ranked according to matching cost shown below. As the search is refined at each level, the difference between accepted and rejected templates decreases

4.3 Detection Examples

The tree-based detection method was tested on real image data. This corresponds to the initialization stage in the hierarchical filter [18, 19]. Figure 6 illustrates the operation of the classifiers at different levels of the tree. In this case the classifiers are based on oriented edges using the chamfer distance and skin colour silhouette. The classifiers at the upper levels correspond to larger regions of parameter space, and are thus less discriminative. As the search proceeds regions are progressively eliminated, resulting in only few final matches. The tree contains 8,748 different templates, corresponding to a pointing hand, restricted to rigid motion in a hemisphere. Figure 7 shows examples of accepted and rejected templates at different tree levels for a different input image.

5 Conclusion

In this paper the concept of a hierarchical cascade of classifiers for locating articulated objects was introduced. The classifiers can be obtained from a geometric

3D model or from training images. The motivation of this research has been ongoing work on hand tracking, in which we seek to combine the merits of efficient detection and tracking. Motivated by the success of tree-based detection, the parameter space is discretized to generate a tree of templates which can be used as classifiers. Even though their performance has been shown to be not as good as classifiers learnt from image data, they have the advantage of being easy to generate and being labelled with a known 3D pose, permitting their use in a model-based tracking framework.

Acknowledgements

The authors gratefully acknowledge the support of Toshiba Research. This work has also been supported in part by EPSRC, the Gottlieb Daimler–and Karl Benz–Foundation, the Gates Cambridge Trust, and the ORS Programme.

References

- [1] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume II, pages 432–439, Madison, USA, June 2003. 106, 107
- [2] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proc. 5th Int. Joint Conf. Artificial Intelligence*, pages 659–663, 1977. 106, 107
- [3] F. C. Crow. Summed-area tables for texture mapping. In *siggraph*, volume 18, pages 207–212, July 1984. 113
- [4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2nd edition, 2001. 109
- [5] P. F. Felzenszwalb. Learning models for object recognition. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume I, pages 56–62, Kauai, USA, December 2001. 108, 109
- [6] D. M. Gavrilu. Pedestrian detection from a moving vehicle. In *Proc. 6th European Conf. on Computer Vision*, volume II, pages 37–49, Dublin, Ireland, June/July 2000. 106, 109
- [7] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to non-linear and non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993. 105
- [8] D. P. Huttenlocher. Monte carlo comparison of distance transform based matching measure. In *ARPA Image Understanding Workshop*, pages 1179–1183, New Orleans, USA, May 1997. 111
- [9] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Analysis and Machine Intell.*, 15(9):850–863, 1993. 109
- [10] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge. Tracking non-rigid objects in complex scenes. In *Proc. 4th Int. Conf. on Computer Vision*, pages 93–101, Berlin, May 1993. 107
- [11] M. Isard and A. Blake. Visual tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. on Computer Vision*, pages 343–356, Cambridge, UK, April 1996. 105

- [12] I. H. Jermyn and H. Ishikawa. Globally optimal regions and boundaries. In *Proc. 7th Int. Conf. on Computer Vision*, volume I, pages 20–27, Corfu, Greece, September 1999. 113
- [13] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proc. 6th European Conf. on Computer Vision*, volume 2, pages 3–19, Dublin, Ireland, June 2000. 107
- [14] C. F. Olson and D. P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *Transactions on Image Processing*, 6(1):103–113, January 1997. 106, 108
- [15] S. Romdhani, P. H. S. Torr, B. Schölkopf, and A. Blake. Computationally efficient face detection. In *Proc. 8th Int. Conf. on Computer Vision*, volume II, pages 695–700, Vancouver, Canada, July 2001. 106
- [16] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume I, pages 746–751, Hilton Head Island, USA, June 2000. 106
- [17] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. 9th Int. Conf. on Computer Vision*, pages 750–757, Nice, France, October 2003. 106
- [18] B. Stenger. *Model-Based Hand Tracking Using A Hierarchical Filter*. PhD thesis, University of Cambridge, Cambridge, U.K., 2004. 109, 111, 112, 114
- [19] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *Proc. 9th Int. Conf. on Computer Vision*, volume II, pages 1063–1070, Nice, France, October 2003. 105, 106, 107, 114
- [20] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla. Learning a kinematic prior for tree-based filtering. In *Proc. British Machine Vision Conference*, volume 2, pages 589–598, Norwich, UK, September 2003. 105
- [21] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *Int. Journal of Computer Vision*, pages 9–19, June 2002. 106
- [22] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume I, pages 511–518, Kauai, USA, December 2001. 106, 113
- [23] O. Williams, A. Blake, and R. Cipolla. A sparse probabilistic learning algorithm for real-time tracking. In *Proc. 9th Int. Conf. on Computer Vision*, volume I, pages 353–360, Nice, France, October 2003. 105