Hand detection using multiple proposals

Arpit Mittal[1]
arpit@robots.ox.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk
Philip H. S. Torr[2]
philiptorr@brookes.ac.uk

[1] Department of Engineering Science, University of Oxford, UK
[2] Department of Computing, Oxford Brookes University, UK

Abstract

We describe a two-stage method for detecting hands and their orientation in unconstrained images. The first stage uses three complementary detectors to propose hand bounding boxes. Each bounding box is then scored by the three detectors independently, and a second stage classifier learnt to compute a final confidence score for the proposals using these features.

We make the following contributions: (i) we add context-based and skin-based proposals to a sliding window shape based detector to increase recall; (ii) we develop a new method of non-maximum suppression based on super-pixels; and (iii) we introduce a fully annotated hand dataset for training and testing.

We show that the hand detector exceeds the state of the art on two public datasets, including the PASCAL VOC 2010 human layout challenge.

1 Introduction

The objective of this work is to detect and localise human hands in still images. This is a tremendously challenging task as hands can be very varied in shape and viewpoint, can be closed or open, can be partially occluded, can have different articulations of the fingers, can be grasping other objects or other hands, etc. Our motivation for this is that having a reliable hand detector facilitates many other tasks in human visual recognition, such as determining human layout [2, 13, 17, 24] and actions from static images [10, 13, 19, 29]. It also benefits human temporal analysis, such as recognizing sign language [5, 14], gestures and activities in video.

Methods based on detecting hands independently using skin detection [27, 28, 30] or Viola & Jones like cascade detectors built from Harr features [23, 25, 26] have had limited success and impact on unconstrained images, perhaps due to lack of sufficient training data. There has been quite some success in detecting hands as part of a human pictorial structure [6, 22, 24] which provides spatial context for the hand position. However, these methods require that several parts of the human (e.g. head and arms) are also visible in the image and have some limitations on the body poses they are trained for (e.g. no self-occlusion).

In this paper we propose a detector using a two-stage hypothesize and classify framework. First, hand hypotheses are proposed from three independent methods: a sliding window hand-shape detector, a context-based detector, and a skin-based detector. Then, the proposals are scored by all three methods and a discriminatively trained model is used to

BMVC 2011 http://dx.doi.org/10.5244/C.25.75

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

Figure 1: **Overview of the method.** (a) Original image. (b) Some of the hypotheses proposed by hand and context detector. Bounding boxes in red are proposed by the hand detector and in green by the context detector. (c) Skin detection and hypotheses generation. (d) Super-pixel segmentation of the image with combined hypothesised bounding boxes from the three proposal schemes. Using super-pixel based non-maximum suppression (NMS), overlapping bounding boxes are suppressed. (e) Final detection after post-processing.

verify them. Figure 1 overviews the detector. The three proposal mechanisms ensure good recall, and the discriminative classification ensures good precision. In addition, we have collected a large dataset of images with ground truth annotations for hands, and this provides sufficient training data for the methods.

We show that this detector can achieve very good recall and precision on unconstrained images from multiple sources, and that it exceeds the state of the art performance of Karlinsky *et al.*[22], and the PASCAL VOC layout challenge[13].

2 Proposal methods

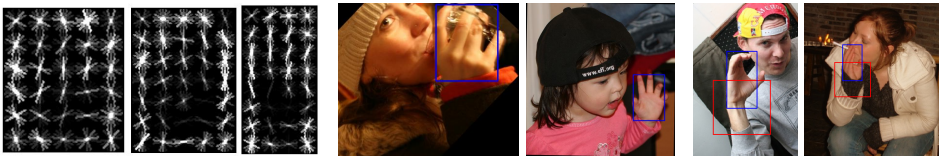
In this section we describe the three proposal methods: shape, context, and skin colour. Each of these delivers a number of hypotheses for a bounding box for the hand, specified as a rotated rectangle (i.e. it is not axis aligned).

2.1 Hand shape detector

This detector proposes bounding boxes for hands using Felzenszwalb *et al.* [16]'s part based deformable model based on HOG features [9]. The detector is a mixture over three components (Figure 2(a)), where each component represent a different aspect of the hand. Learning is done using the training set of the hand dataset (Section 4). The training images are rotated, such that all the hand instances are aligned (as shown in Figure 2(b)). Testing is performed at 36 different rotations of the image (at standard 10° intervals). For each image the proposed bounding boxes are given by the set $B_{HD} = \{b \in B \mid \beta_{HD}(b) > t_h\}$, where β_{HD} is the scoring function [16] of the hand detector, B is the set of all detected hand bounding boxes, and t_h is the threshold of the hand detector chosen to give 90% recall on the training set.

2.2 Context detector

This detector proposes hand bounding box depending on its context. The motivation behind this is that the end of arm may be more visible or recognizable than the hand, and could provide vital cues for hand detection. In order to learn the context, a part based deformable model [16] is trained from the hand bounding box training annotations extended to include



(a) (b) (c)

Figure 2: (a) Root filters for the three components of the hand-shape detector. The first two filters cover frontal pose and the third filter profile. (b) Rotated Training images so that bounding boxes are axis aligned. (c) Context captured around the hand bounding box. The blue rectangle shows the hand bounding box and the red shows the extended box used to capture context around the hand. The context is captured over a region having the same height and twice the width as the hand.

the surrounding region, as shown in figure 2(c). Again, a mixture model with three components is learnt. It should be noted that unlike other methods, which model adjacent body parts such as the arm explicitly, here the area surrounding the hand is instead modelled directly in a discriminative manner. Due to this although the detector is learned over a relatively varied region, it is less altered by occlusion of body parts. For training all the images are rotated so that the bounding boxes have the same orientation (axis aligned) (Figure 2(c)), and testing is performed at 10° intervals of rotation. Hand bounding boxes are obtained from the detected context boxes by shrinking them. Thus, for each image the proposed boxes are given by the set $B_{CD} = \{b \in B \mid \beta_{CD}(b) > t_c\}$, where β_{CD} is the scoring function of the context detector [16], B is the set of all hand bounding boxes, and t_c is the threshold of the context detector chosen to give 90% recall on the training set.

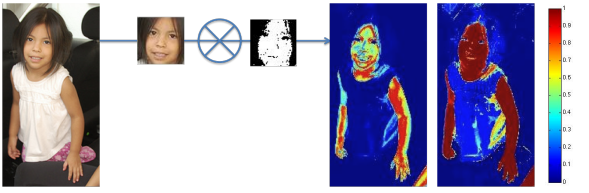
2.3 Skin-based detector

This proceeds in two steps: first skin regions are obtained, then they are used to instantiate hypotheses for hands.

Skin detection. A global skin detector is used to detect the skin regions in the image [8]. The skin detection results are further enhanced by using detected face regions (using the openCV face detector [26]) to determine the colour of skin *in the target image* [18] (Figure 3). This enables skin to be detected despite unusual lighting and colour balances which may be peculiar to the target image. A simple classifier of colour likelihood is used based on a histogram of the face pixels. We proceed in two stages. First, in the manner of hysteresis in the Canny operator, two thresholds (confidences) for likelihood are learnt. Pixels which are above the high threshold are classified as skin. Then pixels which are above the low threshold are also classified as skin if they are spatial neighbours of a pixel above the high threshold (these thresholds are learnt by cross-validation on ground truth segmentations). Second, a bootstrapping stage, the colour of the neighbouring pixels is used to update the colour likelihood classifier and the process is repeated.

Pixels are classified using likelihood here, rather than a posterior or discriminative classifier, as learning a distribution for the background pixels at this stage is error prone due to the possible presence of arms and other people in unknown locations.

Hypothesis generation. Lines are fitted to the skin regions using a Hough transform, and also by finding the medial axis of the blob-shaped skin regions. The medial axis often produces useful lines in the cases where only hands are visible (and the resulting skin regions are then approximately elliptical). The Hough fitting is more useful in the case of skin from arms. The hands are then hypothesised at either ends of the line (Figure 4). The size of



(a) (b) (c) (d)

Figure 3: **Skin detection.** (a) Input image with the face highlighted. (b) A colour histogram is computed from the face skin pixels. (c) Likelihood values for skin pixels. (d) Likelihood values after bootstrapping. Note that many more of the true skin pixels are now detected compared to (c). In the figure, red colour corresponds to the highest value and blue the lowest.

(a) (b) (c) (a) (b) (c)

Figure 4: **Hypotheses generation from the skin regions.** (a) Original image. (b) Lines are fitted to the skin region which are used to hypothesise the extent and orientation of the hand. Hands are hypothesised at both ends of the fitted lines. If the skin region resembles a blob then the whole skin region is hypothesised as a hand. (c) Hand detections remaining after verification using the model.

the box for the hand depends upon the width of the skin region at end of the line. The set of boxes from all lines are the proposals from the skin detection, B_{SD}. No hypothesis is proposed from the facial skin regions. If the face is not detected in the first stage, then skin colour based proposal method is not applied for that image.

3 Hypothesis classification

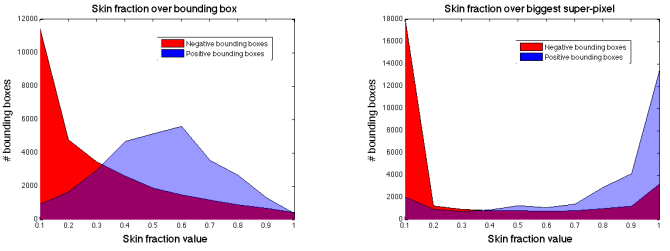
The hypotheses proposed by the different proposal schemes are combined and are then evaluated using a second stage classifier. The complete set of hypotheses is given by the union $B_h = \{B_{HD} \cup B_{CD} \cup B_{SD}\}$. Three scores are then computed for each hypothesised bounding box ($b \in B_h$) as follows:

Hand detector score. A score (α_1) obtained from the hand detector.

$$\alpha_1 = \beta_{HD}(b) \tag{1}$$

where β_{HD} is the scoring function of the hand detector.

Context detector score. In order to include some deformation between the hand and its context, max-pooling of scores is done over all boxes (translated and rotated) having an overlap with the given bounding box, b, above a 0.5 threshold. This gives some degree of invariance to rotation and translation changes. Let $B_h^{'}$ be the set of all bounding boxes having overlap score greater than the threshold value with the given bounding box b. Then the context detector score is given by,



(a) (b)

Figure 5: **Bayes risk plots.** The distribution of skin fraction value for positive bounding boxes is shown in blue, the negative bounding boxes in red, and the Bayes risk in purple. (a) Skin fraction computed over all the pixels of the bounding box. (b). Skin fraction computed for pixels belonging to the biggest super-pixel in the bounding box. It can be seen that (b) has a far lower Bayes risk and therefore can discriminate better between the positive and negative boxes.

$$\alpha_2 = \max_{b_h \in B'_h} (\beta_{CD}(b_h)) \tag{2}$$

where β_{CD} is the scoring function of the context detector.

Skin detection score. For the skin detection score, a straight forward choice would be to use the skin fraction (i.e., the fraction of pixels that are skin in a bounding box b) as the feature. This approach is not suitable for bounding boxes as the boundaries are not tightly aligned with the hand, and may include the arm or other skin regions for example. However, a hand's appearance is often visually coherent and can be obtained as a single super-pixel. Thus, the image is first split into super-pixels [4], then for a bounding box, the skin fraction (α_3) is computed for the biggest super-pixel within it. This gives a better discriminative feature than skin fraction alone (Figure 5).

Classification of scores. The three scores for a given bounding box are combined into a single feature vector, $(\alpha_1, \alpha_2, \alpha_3)$, and a linear SVM classifier [6] is learned over the combined feature space using a standard SVM-solver [20]. This final classifier is used to compute confidence score for all the bounding boxes.

Super pixel based non-maximum suppression. Typically a detection algorithm returns a number of overlapping bounding boxes, and non-maximum boxes are then suppressed depending on their overlap with other high-scoring boxes [7]. However, this suppression in general does not use any visual information from the image. This sometimes results in losing detections for the objects if multiple partially overlapping instances are present in the image.

We propose a modification of the traditional non-maximum suppression (NMS) technique, and instead incorporate further image information into the NMS process. As in the case of the skin detection score, we make use of super-pixels to capture the visual coherence of the hand. The image is first split into super-pixels, and then NMS is applied over all the boxes overlapping the same super-pixel. The overlap threshold for NMS is 0.4. If a box is overlapping more than one super-pixel then it is associated with the one that it is overlapping with the most. Figure 6 shows some examples where super-pixel based NMS performs better than the conventional NMS technique.

To avoid cases where the detected bounding boxes do not fit tightly around the true hand, the NMS surviving box is fitted tightly around its enclosed super-pixels if the super-

(a) (b) (c) (d)

Figure 6: **Comparison of conventional NMS with super-pixel based NMS.** (a) Bounding boxes shown in blue and red are overlapping. (b) Superpixel segmentation of the image. (c) The red bounding box is suppressed by conventional NMS. (d) Super-pixel NMS retains the correct boxes.

pixel resembles a blob. A super-pixel is deemed a blob if the ratio between its major and minor axis is less than a threshold (2.5). Detected hand boxes which overlap with the face regions (localised using the face detector) are also removed as part of the post-processing. By applying these post-processing steps, performance of the system improves significantly (Table 2).

The time taken for the whole detection process is about 2 minutes for an image of size 360×640 pixels on a standard quad-core 2.50 GHz machine. The hand and context detectors employ the efficient cascade implementation of Felzenszwalb *et al.* [15].

4 Hand dataset

We introduce a comprehensive dataset of hand images collected from various different public image data set sources as listed in Table 1. While collecting the data, no restriction was imposed on the pose or visibility of people, nor was any constraint imposed on the environment.

In each image, all the hands that can be perceived clearly by humans are annotated. The annotations consist of a bounding rectangle, which does not have to be axis aligned, oriented with respect to the wrist. Examples are shown in Figure 7.

The data is split into training, validation and test sets in such a way that there is no repetition of any given person among these datasets. Hand instances larger than a fixed area of bounding box (1500 sq. pixels) are used in the subsequent experiments. This gives around 4170 high quality hand instances. The distribution of these images into training, validation and test sets is also given in Table 1. A total of 13050 hand instances are annotated (including the 4170 larger instances). The dataset is available at (http://www.robots.ox.ac.uk/~vgg/data/hands/).

Evaluation Measure. The performance is evaluated using average precision (AP) (the area under the Precision Recall curve). As used in PASCAL VOC [12], a hand detection is considered true or false according to its overlap with the ground-truth bounding box. A box is positive if the overlap score is more than 0.5, where the overlap score (O) between two boxes is defined as: $O = \frac{area(B_g \cap B_d)}{area(B_g \cup B_d)}$, where B_g is the axis aligned bounding rectangle around ground-truth bounding box and B_d is the axis aligned rectangle around detected bounding box.

Training Dataset			Validation Dataset		
Source	**# Ins**	**# Img**	**Source**	**# Ins**	**# Img**
Buffy stickman [1]	438	346	Movie dataset*	649	406
INRIA pedestrian [3]	137	97			
Poselet (H3D) [2]	580	237	**Test Dataset**		
Skin dataset [7]	139	87	**Source**	**# Ins**	**# Img**
PASCAL VOC 2007 train and val set [10]	507	345	PASCAL VOC 2010 human layout val set [13]	98	63
PASCAL VOC 2010 train and val set (except human layout set) [13]	1060	732	PASCAL VOC 2007 test set[10]	562	373
Total number	2861	1844	*Total number*	660	436

Table 1: Distribution of larger hand instances in the hand dataset. '# Ins' is the number of hand instances, and '# Img' the number of images. The movie dataset contains frames from the films 'Four weddings and a funeral', 'Apollo 13', 'About a boy' and 'Forrest Gump'.

Figure 7: Sample images from the hand dataset with bounding box annotations overlaid. In the annotation, rectangle sides are ordered so that the wrist is along the first side marked with '*'.

5 Experimental Results

The model is evaluated on our test dataset and two external datasets (signer dataset [5] and PASCAL VOC 2010 person layout test dataset [13]). We compare to the performance of previous work on these external datasets.

5.1 Hand dataset

For all of the following experiments the model is trained on the hand training dataset and model parameter values are determined on the hand validation dataset, see Table 1.

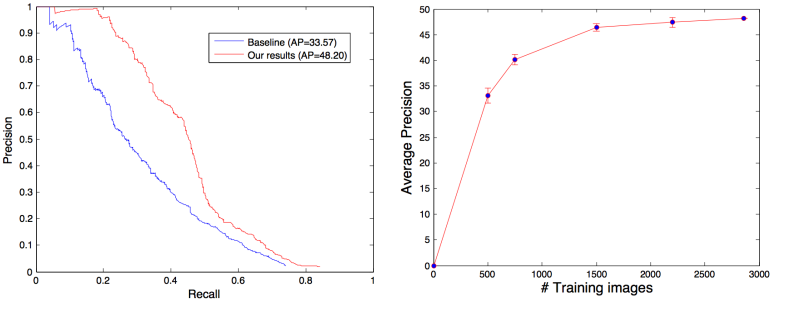
Parameter estimation. The model parameters include: size of context bounding box around the hand, weight values of the second stage classifier and the SVM parameter C.

For the context box, the following parameter values were investigate: ({h, 2w}, {2h, 2w}, {h, 3w} and {2h, 3w}), where 'h' and 'w' are the height and width of the hand bounding box respectively. The AP obtained for theses different sizes is [46.13, 38.75, 44.04, 40.97]. Consequently, a context region is of the same height and twice the width as the hand bounding box is used (refer Figure 2(c)). The weights learnt for the linear SVM used to blend scores from the proposal schemes are $\mathbf{w} = (1, 0.4, 0.36)^\top$ for hand detector score, context detector score and skin detection score respectively. The value of parameter C was learnt as 1.0.

Test set performance. The basic hand detector (i.e. no context or skin detection) is used as the baseline. Table 2 shows the average precision for different proposal schemes after including different scores in the model. The baseline precision is the precision obtained by using just the hand detector as the proposal scheme and the hand score as the only score for the final classifier. Compared to this baseline it can be seen that there is around 15% improvement in average precision and 11% increase in recall by the final model.

Proposal schemes	Hand score	Hand and context scores	Hand, context and skin scores	Recall
Hand	*33.57 / 36.54*	34.23 / 37.78	35.31 / 41.09	74.09
Hand and context	36.19 / 39.22	39.06 / 42.68	41.59 / 47.13	82.12
Hand, context and skin	36.30 / 39.63	39.38 / 43.48	**42.25 / 48.20**	85.30

Table 2: Performance on the hand test dataset in terms of average precision (before post-processing / after post-processing). The final column gives the recall after post-processing. Along a row the variation in performance can be seen for a given proposal scheme after including different scores in the final classifier. The increase in recall with the addition of the different proposal schemes is evident. The baseline is the 'hand proposal, hand score' (in *italics*), and the final performance is shown in bold.



(a) (b)

Figure 8: (a) Precison-Recall curve comparing the baseline and final results. (b) Variation in Average Precision with the number of training instances. To generate the plot, five sets of the specified size are randomly sampled from the hand training set and a model learnt from each. The graph shows the mean AP and standard deviation obtained over the test set for the five models. For the last data-point no such split is done as all of the training data is used. It can be seen that AP increases with the the size of the training set and reaches saturation after 2500 images.

If the conventional NMS is used for post-processing, then the AP of the system reduces from 42.25 to 40.79. The PR curve comparing baseline results with our results after post-processing is shown in the Figure 8(a). Figure 8(b) shows the variation in AP with increase in the training data.

5.2 Signer dataset

The model used for this experiment is trained on the hand training dataset (Table 1). The dataset that is used for this experiment is the '5-signers' dataset which is a collection of frames from five news sequences (39 frames each) with different signers [5].

Karlinsky *et al.* [22] use the ground-truth position and scale of the head bounding box to fit a chain model originating from head up-to the hand. They consider the hand detection to be correct if it is within half face width from the ground-truth location of the hand. They report their detection performance within the top k hand detections per ground-truth hand instance. Table 3(a) compares the result from their method with ours. For evaluation the same criteria is used as that used by the authors.

It can be seen that for $k = 3$ and 4, our method performs better than [22]. For smaller values of k, Karlinsky *et al.*'s method works well because the chain model enables them to disambiguate hands from the background better. However, this can also be a disadvantage of their model as the method requires the position of head at test time, and can not work if the

Setting	1 max	2 max	3 max	4 max
Karlinsky et. al. [22]	84.9	92.8	95.4	96.7
Our method	76.67	90	**95.64**	**97.44**

(a)

Method	BCNPCL	Oxford	Ours
AP	3.3	10.4	**23.18**

(b)

Table 3: (a) Comparison of results on the Signer dataset.'1 max', '2 max' etc. are the detection performance within the top 'k' hand detections per ground-truth hand instance. (b) Comparison of our method with other submissions for PASCAL VOC 2010 person layout challenge for hand detection task [13]. Scores are obtained by submitting results to the competition evaluation server.

Figure 9: Examples of high-scoring detections on the three datasets. Top row: images from the hand dataset; Middle row: images from the Signer dataset; Bottom row: images from the PASCAL VOC 2010 person layout test set.

head or some-other body parts are occluded. Our method does not require other body parts to be visible (and is therefore not restricted to images having un-occluded humans in frontal poses).

5.3 PASCAL VOC 2010 person layout test dataset

The dataset for PASCAL VOC 2010 person layout challenge [13] has 320 images with 505 humans annotated. Bounding boxes are provided around every human figure and the task is judged by how well the three body parts (head, hand and feet) are predicted *individually* in the given region. The prediction of a part is considered to be correct if the overlap score with the ground truth is more than 0.5. A method must provide a single score for each human annotated, and results are ranked according to this score (i.e. head, hand and feet detections are not ranked separately). The performance is measured using AP.

For this problem, we train the model on hand training dataset (Table 1) and evaluate it for competition 8 (i.e., training on own data). Each of the provided human bounding boxes is re-sized such that the minimum width is at-least 300 pixels. This is done to ensure that the hands in the image are reasonably large for detection. For every figure, either one or two hand bounding boxes are returned depending upon the confidence scores obtained from the model. The final confidence score for the human is the average of the scores for each of the hands.

As shown in table 3(b), we report a very good performance over this dataset, beating our previous winning entry by a factor of two.

A sample of high scoring detections from the three datasets are shown in figure 9.

6 Conclusions and future work

We have demonstrated that the proposed two stage hypothesise and classify method is capable of improving recall and precision over state of the art results. A natural extension would be to add a fast first stage to the cascade so that the time for the entire process can be reduced.

Acknowledgements. We are grateful for financial support from the ERC grant VisRec no. 228180, and ONR MURI N00014-07-1-0182.

References

[1] http://www.robots.ox.ac.uk/∼vgg/data/stickmen/index.html.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, 2009.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Proc. CVPR*, 2009.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IJCV*, 2009.

[5] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC.*, 2008.

[6] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[7] J. F. Canny. A computational approach to edge detection. *IEEE PAMI*, 8(6):679–698, 1986.

[8] C. Connaire, N. O'Connor, and A. Smeaton. Detector adaptation by maximising agreement between independent data sources. In *OTCBVS'07 - IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2007.

[9] N. Dalal and B Triggs. Histogram of Oriented Gradients for Human Detection. In *Proc. CVPR*, volume 2, pages 886–893, 2005.

[10] V. Delaitre, I Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proc. BMVC.*, 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, Jun 2010.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[14] A. Farhadi and D. Forsyth. Aligning ASL for statistical translation using a discriminative word model. In *Proc. CVPR*, pages 1471–1476, 2006.

[15] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proc. CVPR*, pages 2241–2248, 2010.

[16] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010.

[17] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, Jun 2008.

[18] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *IEEE Workshop on Robot and Human Interactive Communication*, 2002.

[19] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 2009.

[20] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[21] M. J. Jones and J.M. Rehg. Statistical color models with application to skin detection. *IJCV*, 2002.

[22] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *Proc. CVPR*, 2010.

[23] M. Kolsch and M. Turk. Robust hand detection. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2004.

[24] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Efficient discriminative learning of parts-based models. In *Proc. ICCV*, 2009.

[25] E.J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2004.

[26] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001.

[27] Y. Wu and T. S. Huang. View-independent recognition of hand postures. In *Proc. CVPR*, 2000.

[28] Y. Wu, Q. Liu, and T. S. Huang. An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization. In *Proc. Asian Conf. on Computer Vision*, 2000.

[29] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *Proc. CVPR*, 2010.

[30] X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2000.