Learning and detection of object landmarks in canonical object space

Joni-Kristian Kamarainen*, Jarmo Ilonen
Machine Vision and Pattern Recognition Laboratory
**Computational Vision Group, Kouvola*
Lappeenranta University of Technology

Abstract

This work contributes to part-based object detection and recognition by introducing an enhanced method for local part detection. The method is based on complex-valued multiresolution Gabor features and their ranking using multiple hypothesis testing. In the present work, our main contribution is the introduction of a canonical object space, where objects are represented in their "expected pose and visual appearance". The canonical space circumvents the problem of geometric image normalisation prior to feature extraction. In addition, we define a compact set of Gabor filter parameters, from where the optimal values can be easily devised. These enhancements make our method an attractive landmark detector for part-based object detection and recognition methods.

1. Introduction

In visual object categorisation (VOC) one wishes to identify all objects in an image and in object detection to locate them more accurately by a bounding box or object landmarks. Most VOC methods utilise the popular bag-of-features (BoF) approach [7, 20, 16], but part-based methods have certain advantageous properties [6, 9]. One disadvantage of part-based methods is the need of object region or landmark annotation. The extra annotation work is justifiable for certain applications, since the part-based approach often provides superior localisation accuracy [12]. The overall accuracy is limited by the accuracy of local part detector.

In our previous work we developed a particularly accurate and robust method for landmark detection [13]. One weakness in our method was the geometric normalisation of training image ensemble, which was assumed given. In this work, we propose a canonical object space, which circumvents the normalisation problem. The canonical space is estimated from the annotated landmarks and therefore images can be automatically normalised by registering them to the canonical space. The canonical space estimation algorithm is similar to the alignment method for point patterns in [5], with exceptions that patterns are iteratively (not batch training) transformed to an updated mean shape space, and in addition, the alignment is not restricted to approximate similarity, but can be any 2D homography. The detection method augmented with the canonical object space normalisation step is illustrated in Fig. 1.

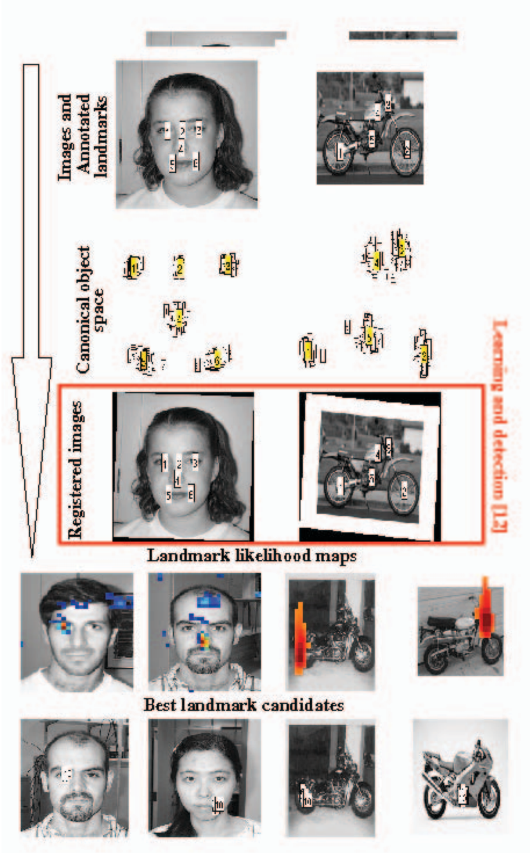

Figure 1. Two Caltech-101 categories, their canonical object spaces and detection results (colour image).

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.348

1.1. Related work

The part-based approach to object detection and recognition was proposed by Fischler and Elschlager in 1973 [11]. The approach contains two intrinsic tasks: 1) description and detection of local parts (landmarks) and 2) search over spatial configurations of the detected landmarks (constellation search). Felzenszwalb and Huttenlockher have proposed a statistical constellation model which copes with part articulation (e.g. human limbs) [9]. For landmarks, their method utilises iconic representations constructed from Gaussian derivative filters [17]. For the constellation search, Carneiro and Jepson proposed a geometric prediction of local parts [4]. Parts in their method are represented by log-polar space histograms [3]. The authors have proposed a similar geometric transformation constellation search formulated in a statistical framework [14] and utilising our learnable detectors [13]. The BoF methods typically utilise interest point detectors and descriptors instead of fixed local parts, but several particularly successful hybrids of interest points and constellation search have been proposed [1, 2].

The proposed canonical object space which enhances our detector is similar to the shape model by Cootes et al. [5]. In their original work edge detectors were used, but they have recently changed these to more general features similar to ours [6].

2 2D object landmarks

2D object landmarks denote some local structures which appear visually similar in different instances of the same category. We have not investigated the landmark selection itself, but accept any manually annotated parts, such as the six face and five motorbike landmarks illustrated at the top of Fig. 1. Our detector method is the same as in [13], but as novel contributions we automate its image normalisation requirement by the proposed canonical object space in Sec. 2.1 and revise the descriptor part in Sec. 2.2. For completeness we also describe the learning and detection in Sec. 2.3.

2.1. Canonical object space

In [13] training images were assumed aligned for normalised object appearance. For the most data sets this is not the case. We circumvent the problem by the canonical object space, where the geometric variations, such as translation, rotation, scaling and skew, are minimised. The canonical space alignment can be performed automatically since annotated landmarks allow 2D homography estimation using point correspondence. One of the images is randomly selected as a seed and the remaining images are iteratively mapped to it and the space simultaneously optimised. 2D homography can be restricted to isometry, similarity, affinity or full homography, which can be efficiently computed by the linear methods: Umeyama for isometry and similarity (requiring at least 2 correspondence) [19], a restricted direct linear transform (DLT) for affinity (3 correspondence) and the normalised DLT for projectivity (4 correspondence) [15]. The code is available at http://www.it.lut.fi/project/homogr.

Our approach is based on the mean shape model by Cootes et al. [5]. We utilise their iterative method, but replace the approximate similarity transformation by the linear homography algorithms. The method iterates through training examples and outputs the final canonical object space, where the images are registered by a standard 2D transformation with bicubic interpolation. Another advantage is that if the training and test sets are well in correspondence, then the canonical space is a good estimate of the optimal pose for the landmark detection. The pseudo code for the procedure is given in Alg. 1 and the estimated canonical object spaces for 798 motorbike images are illustrated in Fig. 2. The original images are clearly misaligned, but the similarity restricted canonical space provides very good alignment for all images.

Algorithm 1 Canonical object space.

```
Select a random seed image and use its landmarks as the
initial object space.
**for all** images **do**
    Estimate homography to the object space.
    Transform image landmarks to the object space.
    Refine the object space by taking the average of all
    transformed landmarks.
**end for**
Return the final canonical object space and all images and
their landmarks projected to the canonical space.
```

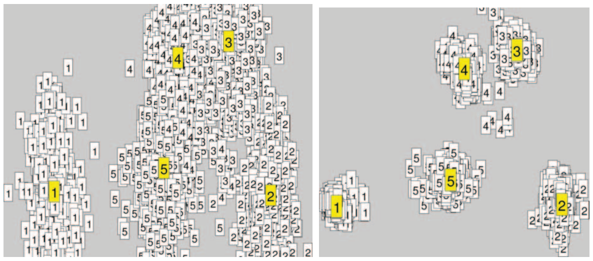

Figure 2. Canonical object space (yellow labels) for Caltech-101 Motorbikes with all images aligned: original (left) and similarity (right).

2.2. Low level features

Our low-level feature is a complex vector formed from responses of a Gabor filter bank [13]. The feature establishes only moderate geometrical invariance, but we have introduced simple shift operations for fast search over scale and rotation. Therefore, by using a fast classifier to learn landmarks from their Gabor responses in the canonical object space, we can establish an efficient detector which outputs best landmark candidates invariant to translation, scaling and rotation.

The Gabor bank is defined by the following parameters: the base frequency f_{max}, frequency scaling factor k, number of frequencies m and number of orientations n. There are no rules for selecting optimal parameters, but based on our earlier experience we define an empirical set of possible values: $m = \{3, 4, 6\}$, $n = \{4, 6, 8\}$, $k = \{\sqrt{2}, \sqrt{3}, \sqrt{4}\}$, and $f_{max} = \{1/20, 1/30, 1/40, 1/50\}$. Good results have been achieved by optimising the combination of these values. The values are valid for VGA size images, but worked well for the varying size images in the Caltech-101 data set [8]. For non-VGA size images the frequencies may need proportional adjustments as they have been defined in pixels. The detector code is available at http://www.it.lut.fi/project/simplegabor.

2.3. Learning and detection

Our main objective is to detect N best candidates of each landmark. It turns out that the class likelihood value is the correct statistical tool to rank the features. For computing the landmark likelihoods we need to estimate their probability density functions (pdfs). Gaussian mixture models are the baseline method for pdf estimation and this can be performed unsupervisely with the Figueiredo-Jain algorithm [10]. At first, the Gabor bank responses are computed at each landmark location in the canonical space for training images. Secondly, landmark pdfs are estimated from low level features using the Figuiredo-Jain algorithm. Our F-J GMM implementation is available at http://www.it.lut.fi/project/gmmbayes.

The detector output is a map of the best likelihoods for each landmark over rotations and scales. By selecting the highest likelihoods the best landmark candidates are found. The likelihood maps for two face landmarks (left eye and tip of nose) and two motorbike landmarks (rear tyre and front joint) are demonstrated in Fig. 1. The image show only the best 5% of the values. In the same image also 10 best landmark candidates are shown for different images (left eye and right mouth corner for faces and rear tyre and motor for motorbikes).

3. Experiments

To demonstrate the performance of our method in object landmark detection we utilised the popular Caltech-101 data set for benchmarking visual object categorisation methods. We selected two visually different categories for which manually selected landmarks were annotated: Faces_easy (tot. 435 images) and Motorbikes (tot. 798 images). We randomly divided images into equal size training and testing sets. The optimal Gabor bank parameters were set using the cross-validation: $f_{max} = 1/20, k = \sqrt{3}, m = 4, n = 6$ for Faces_easy and $f_{max} = 1/30$, $k = \sqrt{2}$, $m = 5$, $n = 6$ for Motorbikes. No remarkable differences in the results occurred even if the same parameters were used for the both. Similarity transformation was used for the canonical object space. No special attention was paid for the landmark selection and it turned out that our face landmarks were very good, but for the Motorbikes certain landmarks were unreliable. That can be explained by the significant differences between scooters, road bikes, enduros, etc.

In order to quantitatively evaluate the detection and localisation accuracies we adopted the d_eye measure used in the face detection evaluation [18]. In the original evaluation protocol the two eye centres are used to normalise distances for images of different size. Respectively, we selected the first and last landmarks for the normalisation. The following three normalised distance thresholds were used to measure the accuracy: 0.05, 0.10 and 0.25 (see Fig. 3 for illustrations).

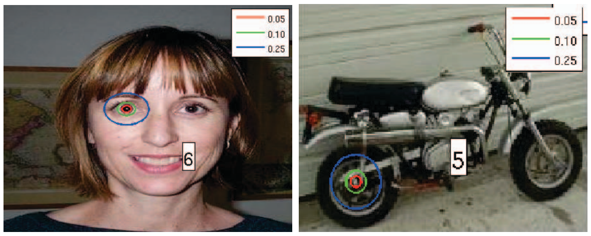

Figure 3. Accuracy thresholds for successful detection (0.05: red, 0.10: green and 0.25: blue). The corresponding circles are drawn on the first landmark and the last denoted by its label.

The detection results for the 100 best landmarks are shown in Fig. 4. The cumulative matching histograms illustrate the total proportion of the correct landmarks as the functions of the total number of extracted landmarks. The detection bars illustrate the same information image-wise: how many correct landmarks are detected in different images. The detection was almost perfect for Faces_easy since nearly 100% of all landmarks were correctly detected within the tightest threshold within only 10 best candidates. The Motorbikes category was more difficult and the corresponding detection accuracies were 65% for the tightest threshold and 90% for the loosest.

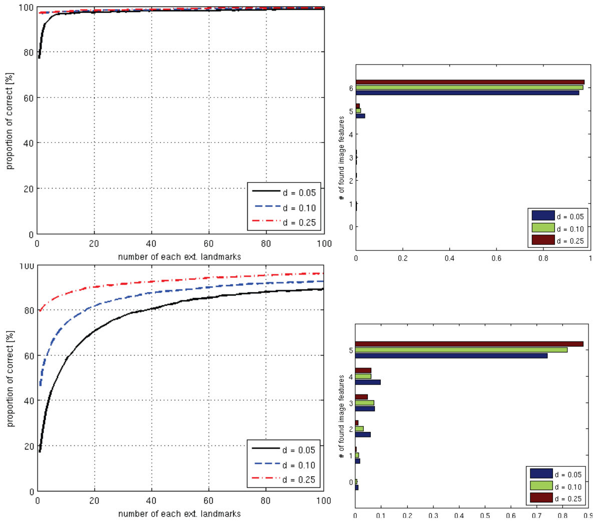

Figure 4. Detection results (Faces_easy top and Motorbikes bottom): cumulative match histograms (left) and detection bars (right).

4. Conclusions

We have contributed to part-based object detection and recognition by improving the previously defined method for object landmark detection. We proposed the canonical object space for aligning images prior to feature extraction and landmark learning. We also proposed an automatic method for the canonical space construction. These enhancements make our method an attractive local part detector for part-based methods.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE PAMI*, 26(11), 2004.

[2] A. Bar-Hillel and D. Weinshall. Efficient learning of relational object class models. *Int J Comput Vis*, 77:175–198, 2008.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE PAMI*, 24(24), 2002.

[4] G. Carneiro and A. Jepson. Flexible spatial configuration of local image features. *IEEE PAMI*, 29(12), 2007.

[5] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1), 1995.

[6] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41:3054–3067, 2008.

[7] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV Workshop*, 2004.

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 2004.

[9] P. Felzenszwalb and D. Huttenlockher. Pictorial structures for object recognition. *Int. J. of Computer Vision*, 61(1), 2005.

[10] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *IEEE PAMI*, 24(3):381–396, 2002.

[11] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computers*, 22(1):67–92, 1973.

[12] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas. Feature-based affine-invariant localization of faces. *IEEE PAMI*, 27(9):1490–1495, September 2005.

[13] J. Ilonen, J.-K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kälviäinen. Image feature localization by multiple hypothesis testing of Gabor features. *IEEE Trans. on Image Processing*, 17(3):311–325, 2008.

[14] J.-K. Kamarainen, M. Hamouz, J. Kittler, P. Paalanen, J. Ilonen, and A. Drobchenko. Object localisation using generative probability model for spatial constellation and local image features. In *ICCV Workshop*, 2007.

[15] J.-K. Kamarainen and P. Paalanen. Experimental study on fast 2d homography estimation from a few point correspondences. Research report 111, Department of Information Technology, Lappeenranta University of Technology, 2009.

[16] M. Marszałek and C. Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, 2008.

[17] R. Rao and D. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505, 1995.

[18] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24:882–893, 2006.

[19] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE PAMI*, 13(4):376–380, 1991.

[20] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.