Local Feature Based Unsupervised Alignment of Object Class Images

Jukka Lankinen
http://personal.lut.fi/users/Jukka.Lankinen
Joni-Kristian Kamarainen
http://personal.lut.fi/users/joni.kamarainen

Machine Vision and Pattern
Recognition Laboratory
Lappeenranta University of Technology,
Kouvola Unit, Finland

Abstract

Alignment of objects is a predominant problem in part-based methods for visual object categorisation (VOC). These methods should learn the parts and their spatial variation, which is difficult for objects in arbitrary poses. A straightforward solution is to annotate images with a set of "object landmarks", but due to laborious manual annotation, semi-supervised methods requiring only a set of images and class labels are preferred. Recent state-of-the-art VOC methods utilise various approaches to align objects or otherwise compensate their geometric variation, but no explicit solution to the alignment problem with quantitative results can be found.

The problem has been studied in the recent works related to "image congealing". The congealing methods, however, are based on image-based processing, and thus require moderate initial alignment and are sensitive to intra-class variation and background clutter. In this work, we define a local feature based algorithm to rigidly align object class images. Our algorithm is based on the standard VOC tools: local feature detectors and descriptors, correspondence based homography estimation, and random sample consensus (RANSAC) based spatial validation of local features. We first demonstrate how an intuitive feature matching approach works for simple classes, but fails for more complex ones. This is solved by a spatial scoring procedure which is the core element in the proposed method. Our method is compared to a state-of-the-art congealing method with realistic and difficult Caltech-101 and randomised Caltech-101 (r-Caltech-101) categories for which our method achieves clearly superior performance.

1 Introduction

In the baseline approach to visual object categorisation (VOC), Bag-of-Features (BoF), the classes are learned from automatically detected local features, but their location information is omitted (*e.g.*, [7]). BoF methods succeed thanks to effective discriminative learning algorithms, but state of the art performances for the most challenging benchmark data sets have been achieved by utilising both detected local features and their spatial configuration, "constellation". These methods are referred to as the *part-based approach* to VOC. With certain techniques, such as multi-resolution processing [18] or special projections [3], some spatial information can be incorporated in the BoF framework, but these require non-intuitive ad hoc processing to gain robustness to standard geometric variations, such as rotation, scaling and translation. Due to its efficiency, discriminative learning has also been applied for the part-based approach [17], but it suffers from limited geometric invariance. The most popular part-based solutions seem to utilise generative model for the spatial constellation of the local parts. An extra challenge in that case is, that inferring a generative model is difficult if the training objects are not aligned but in unknown poses.

The first part-based VOC methods required annotated object landmarks to transform images into a "canonical object space", where geometric variation was removed [21]. It is laborious to select and annotate hundreds of images with good landmarks, and therefore methods requiring less supervision have attracted attention. The most popular approach are *semi-supervised* methods which automatically select and learn the local parts and the constellation model from a set of images containing examples of a same class. That means that the unsupervised image alignment must occur at some point of the algorithms. Effective semi-supervised methods exist [1, 2, 4, 14, 16, 29], but none of the works explicitly define an alignment method nor report quantitative performance of the alignment or its effect to the overall performance.

Outside the "VOC community", the unsupervised alignment has been recognised as its own problem referred to as "spatial image congealing" [19] and a number of congealing methods have been proposed [5, 6, 11, 13, 19, 32]. These methods are mainly seminal work to Learned-Miller [19, 26] extending and improving the original algorithm. The main drawback of these congealing methods is that they are iterative optimisation methods working on pixel-level and thus require at least moderate initial alignment in order to converge properly. The original application areas of the congealing were different, but recently Huang *et al.* [13] proposed congealing for the alignment of natural objects.

Our work deviates from the congealing works by the fact that we utilise local features instead of the pixel level processing, i.e. feature-based congealing. Our solution is more similar to those used in the part-based VOC methods, but we explicitly define the alignment algorithm and measure its performance. We report both qualitative (average images with and w/o the alignment) and quantitative (alignment errors of manually annotated landmarks) performance for difficult categories in the publicly available Caltech-101 data set and its randomised version: r-Caltech-101. We make the following important contributions:

- Propose a feature-based alignment algorithm for unsupervised alignment of object class images.
- Report qualitative and quantitative alignment performance for realistic and difficult categories in Caltech-101 and its randomised version, r-Caltech-101, and compare our results to the state-of-the-art congealing method [13].

All source code and data will be made publicly available.

1.1 Related work

Our feature-based algorithm is fundamentally different to the congealing methods [5, 6, 10, 11, 13, 19, 26, 32] which are iterative optimisation methods using pixel-level differences. Similar ideas have been used in the recent semi-supervised part-based VOC methods [1, 2, 4, 14, 16, 27, 29]. The main weakness in these works is that they do not explicitly define or study the alignment problem.

Belongie *et al.* [1] study alignment of shapes and silhouettes and their method is not directly applicable to natural images of object classes. The same holds for the work of Berg *et al.* [2], which extends the previous method by using edge detectors to select putative candidates and geometric blur as the feature descriptor. Pair-wise non-rigid transformations are searched by minimising a cost function including a transformation model parameters and descriptors. They reported classification performance for a task where a shortlist of manually segmented exemplars were used and whether the method can be used for the unsupervised alignment is unclear.

The recent methods by Chen *et al.* [4] and Jiang *et al.* [14] need manually marked bounding boxes for the objects and the method by Kokkinos and Yuille [16] annotated features for the initial training stage. These methods require less supervision than the full manual landmark selection and annotation, but do not provide unsupervised alignment. Kokkinos and Yuille also tried their method with automatically selected landmarks, but with significantly worse results.

The most similar works to ours are the VOC methods by Philbin *et al.* [27] and by Todorovic and Ahuja [29]. The method by Todorovic and Ahuja forms a tree structure of nested local parts and can be used for segmentation, but is not directly applicable for the alignment. The Philbin *et al.* method first performs BoF search to find a shortlist of class candidates and then performs a spatial scoring to select the most likely match based on local feature similarities and the number of corresponding features under homography transformation. Their method, however, uses local part matching directly, and therefore works best for searching a specific object.

1.2 Method overview and restrictions

Our assumptions are general and shared by many part-based VOC methods. Most importantly, we assume that the categories can be represented by class specific local features, whose saliency triggers local feature detectors and which can be matched using local feature descriptors. That forms the basis for the feature scoring algorithm in Sec. 2. The category specific local features, object landmarks, can also be used to spatially align the images by geometric transformations estimated from their point correspondences. That forms the basis for the spatial scoring algorithm in Sec. 3. We restrict the transformation type to rigid 2D transformations. Fig. 1 demonstrates the ideal accuracy of affine transformation to perform rigid alignment of various Caltech-101 classes.

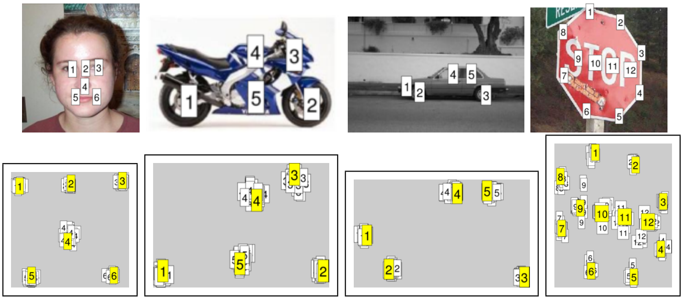

Figure 1: Caltech-101 images with annotated landmarks. Bottom: all images projected to the first set (denoted by the yellow tags). The two standard deviations of the image diagonal normalised projection errors are: 0.0158, 0.0297, 0.0194 and 0.0460, respectively.

We use the existing local feature detection and description methods. Detected local features are not applicable as such for the alignment since false matches occur frequently and therefore scoring best matches leads to bad object landmarks as demonstrated in Sec. 2. The feature scoring is based on pair-wise matching matrices which are also used by the proposed method (Sec. 3) to select putative matches which are verified spatially. Our method utilises a "seed image" where the remaining images are registered for joint-alignment. Our algorithm scores K best seed features over the whole image ensemble. It is the spatial scoring algorithm which makes our method robust to appearance variation, pose variance and background clutter. The main bottleneck is the seed selection, which can be automatically done by introducing some scoring method, or by simply testing a set of random seeds and manually selecting the one which produces the best average image.

2 Unsupervised alignment using feature scoring

Feature scoring provides an intuitive approach for selecting the best object landmarks from a seed image. Scores for K best matching seed features are accumulated over an image ensemble, and features with the highest scores are selected as the object landmarks. The other images are then aligned to the seed using only the best landmarks. This approach is similar to stereo and wide baseline matching [28, 30], except that they do not need the scoring stage or the landmark selection. The scoring, however, improves robustness, which is needed since we want to select those features which frequently appear in the image set, and since appearance variation between class examples can be huge. That is not the case in stereo and wide baseline matching where a scene remains the same.

2.1 Local feature extraction

For the local feature extraction step we adopt one of the most successful detectors, Hessian-affine [22], and descriptors, SIFT [20]. The combination of these performed well in the comparisons of interest point detectors [25] and descriptors [23], in the BoF categorisation experiments [24, 33], and in our preliminary tests as well.

2.2 Feature scoring

The basic idea is to select the seed features which match well to the similarly detected features in other images. Overall performance can be computed using, for example, the SIFT descriptor distances. The simplest solution is to sum the seed feature distances and select the ones with the K smallest sums. That solution, however, is not robust to missing or occluded landmarks and a single image may have an undesirably strong effect to the sum. Therefore, we replace the sum with a ranking-based accumulation: our algorithm accumulates scores of the K best matches per image. Hypothetically, the best landmarks should finally appear as the top scoring seed features. The algorithm is given in Algorithm 1.

Examples of the best landmarks are shown in Fig. 2. Some of them are good and locate in the object areas, but many of them are on rich texture locations in the background or object edges. In general, the feature scoring method in Alg. 1 can find good *object specific landmarks*, such as for face images of a same person, but not good class specific landmarks. The only useful part are the feature score matrices, $S_{N \times M_i}$, between a seed and other images i. These are used by the spatial scoring procedure described next.

Algorithm 1 Landmark selection by feature scoring.

```
Select the seed image and remove it from the image set.
Extract seed interest points and form their descriptors (tot. of $N$).
Initialise score vector $s_N$for $N$ candidates.
for all images (indexed with $i$) do
    Extract interest points and form their descriptors (tot. of $M_i$).
    Compute distance from each seed point to each image point: feature score matrix $S_{N \times M_i}$.
    Increment scores for the $K$ best matching seed features in $s_N$.
end for
Return coordinates and descriptors of the $K$ best scoring seed interest points.
```

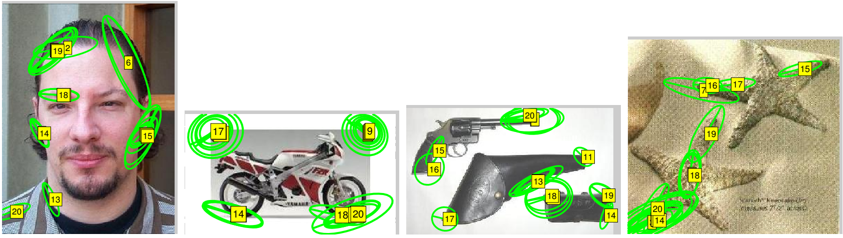

Figure 2: Twenty best landmarks found for different categories using Algorithm 1.

3 Unsupervised alignment using spatial scoring

The main problem in the feature scoring is that it does not use spatial information of the detected features. The spatial information can be used by scoring the points which match under a selected transformation, such as 2D homography. For point correspondences homography transformation can be computed using the standard linear methods: Umeyama [31] for isometry and similarity, and a restricted version of the direct linear transform (DLT) for affinity (the standard DLT accounts for projectivity [12]). Over R iterations we randomly select the minimum number of correspondences (two for isometry and similarity, and three for affinity), estimate homography, transform image points to the seed and accumulate scores of the features matching within a pre-set distance limit τ. Justification to this random approach comes from the random sample consensus (RANSAC) robust estimation [9] and is also used in stereo and wide baseline matching, except that we do not seek for a single solution but accumulate scores over a number of random iterations and number of images. Our algorithm is also different from Philbin *et al.* [27] who use the best matching features over all possible combinations to verify that at least a fixed number of spatially matching features can be found, and then again use feature descriptors of the verified features for classification. Our spatial scoring method for the landmark selection is given in Algorithm 2 and its main computational factor is the number of random iterations R. The other parameters are the match threshold τ and the number of best matches $L \in [1, M_i]$ included from the feature distance matrices $S_{N \times M_i}$.

The spatial scoring algorithm outputs the best K landmarks based on the top scores. The top scoring seed features represent landmarks which have been independently verified by other features in a similar configuration in other images. The spatial scoring landmarks for various Caltech-101 categories are illustrated in Fig. 3. Now the landmarks for even the most

Algorithm 2 Landmark selection by spatial scoring.

```
Select the seed image and remove it from the image set.
Extract seed interest points and form their descriptors (tot. of $N$).
Initialise score vector $s_N$ for $N$ candidates.
for all images (indexed with $i$) do
    Compute the feature distance matrix $S_{N \times M_i}$. // Similar to Alg. 1
    Initialise image-wise score vector $v \leftarrow 0$.
    for $R$ random iterations do
        Select random seed features (minimum number required for the selected homography).
        Select random correspondences from the $L$ best matches in $S_{N \times M_i}$.
        Estimate homography from the seed to the image features.
        Transform all image features to the seed space.
        for all seed features excluding the selected (indexed with $j$) do
            if the seed feature has a match closer than $\tau$ within the $L$ best in $S_{N \times M_i}$ then
                Increment the score for that seed feature: $v(j) \leftarrow v(j) + 1$.
            end if
        end for
    end for
    Sort the scores in $v$ and increment the $K$ highest seed scores in $s_N$. // Note, that each image has
    now equal contribution.
end for
Return coordinates and descriptors of the $K$ best scoring seed interest points.
```

complex categories are within the object area and thus represent the object class much better than in Fig. 2.

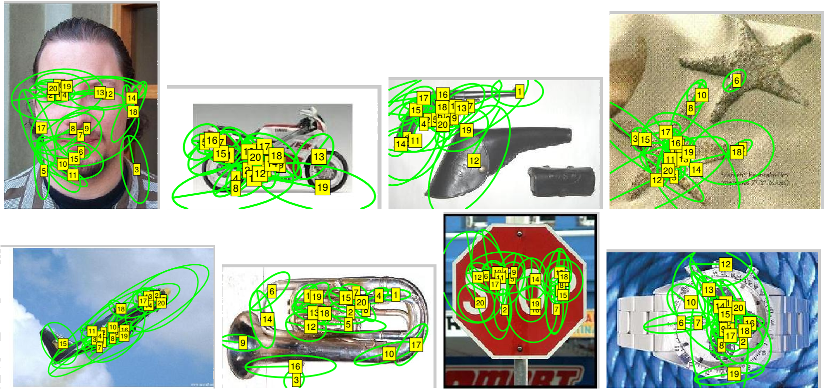

Figure 3: Twenty best landmarks found for different categories using Algorithm 2.

3.1 Alignment using the selected seed landmarks

With the automatically selected object landmarks the alignment procedure itself is straightforward. For a number of random iterations, the minimum number of seed landmarks are selected, then homography is estimated to randomly selected correspondence points within the best matches, and finally, the homography which produces the highest number of inliers is selected and the transformation re-estimated using all inliers. This stage is very similar to stereo and baseline matching, except that again not only the best but a few best matches should be used for robustness.

4 Performance evaluation

The VOC methods reviewed in Sec. 1.1 ([1, 2, 4, 14, 16, 27, 29]) did not report alignment results but only the VOC classification performance. The congealing methods reviewed in the same section used simple classes, such as binary images of digits, and did not provide quantitative results but only the average images before and after congealing. The average images are good for visualisation, but not for a reliable method comparison. Cox *et al.* [5] exemplified their method with handwritten digits in the MNIST data set and face images in the MultiPIE data set. As an exception, they also reported quantitative results as the cumulative distribution of the RMS point error to manually annotated landmarks of the MNIST faces. We report both the average images and the cumulative RMS error curves.

In our experiments, we used by an order of magnitude more difficult images as compared to the congealing works: object class examples in Caltech-101 [8], which is the most popular VOC benchmark. Caltech-101 contains object classes with natural visual appearance variation and with varying background. The main problem of Caltech-101 is that the objects are mainly in the image centre and pose variation is very limited. Moreover, some classes have virtually no background or the background remains the same. These problems make the data set bad for comparing alignment methods and for this reason, we also report the results for the recently published randomised Caltech-101 (r-Caltech-101) [15]. In r-Caltech-101 the backgrounds have been replaced with random Google landscape images and the objects transformed to random poses (scale, translation, rotation). To compute the cumulative RMS curves, we annotated all test images with 5-12 landmarks. The selected parameters for our method in Alg. 2 were: R=1000 random iterations, $\tau = 0.05$ and $K = L = 10$.

The RMS performance represents the proportion of the images (y-axis) for which a specific accuracy (normalised distance: x-axis) is achieved. The distance normalisation was performed by dividing pixel errors with the seed image diagonal which makes the error measure resolution independent. The "ideal result" was computed using the groundtruth landmarks for the alignment. The results are reported for the following four categories: *Faces*, *Motorbikes*, *cars_side*, and *stop_signs*, which represent small (stop signs), moderate (Faces) and large (motorbikes and car side views) visual appearance variation.

In Fig. 4 are shown the average images for the selected classes in r-Caltech-101 without alignment, using a state-of-the-art method by Huang *et al.* [13] and for our method with the best seed. Clearly, our method provides a recognisable average classes while the state-of-the-art congealing suffers from the fact that the initial alignment is poor and backgrounds cluttered. The error graphs are shown in Fig. 5 for the corresponding images in Fig. 4. In addition, we have added the results with the same images in the original Caltech-101 in order to verify that similar results were achieved with the original data. It is noteworthy, that the cumulative error of 0.2 is very large and the most important region is between 0.0 and 0.1. For example, the two standard deviations in Fig. 1 (covering 95% of all errors) were all below 0.05, which was used as the matching threshold τ. The differences of our method between Caltech-101 and r-Caltech-101 result graphs can be explained by the fact that the best seed can be different for Caltech-101. Our method, however, succeeds with Caltech-101 and r-Caltech-101 almost equally on average.

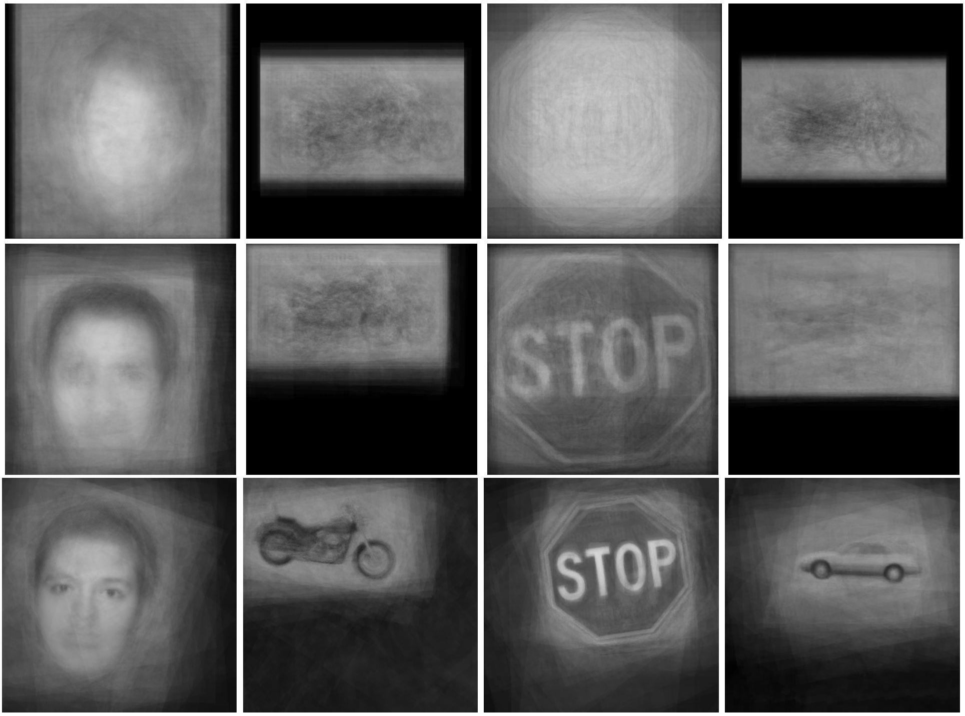

Figure 4: Qualitative alignment results for r-Caltech-101: unaligned average (top), Huang *et al*. [13] (middle), and our method (bottom).

5 Discussion and conclusions

This paper investigated unsupervised object class image alignment, which is a predominant sub-problem in part-based methods for semi-supervised visual object categorisation. Similar methods to ours have been used in the recent VOC methods, but to the authors' best knowledge this work is the first to explicitly investigating the problem, providing an algorithm, and reporting qualitative and quantitative alignment performance.

The alignment problem has been studied in the context of "image congealing", but the congealing methods are not robust to drastic misalignment and background clutter. In the experimental part of our work, our method achieved superior results to a state-of-the-art congealing method.

The main limitation of our method is the seed selection. Automatic seed selection will be addressed in our future work, but as a straightforward solution in this work, we computed the average alignment images for every seed and manually selected the visually most plausible. This simple solution does not violate the semi-supervised principle where input images need to be processed anyway for providing class labels. However, in our case no annotation for landmarks, boundary boxes, or regions is needed.

Future work shall further try to combine our method with the state-of-the-art part-based VOC methods, using our method at the core of their algorithms for image alignment and part selection. Note, that as its side result, our method automatically provides also a set of the

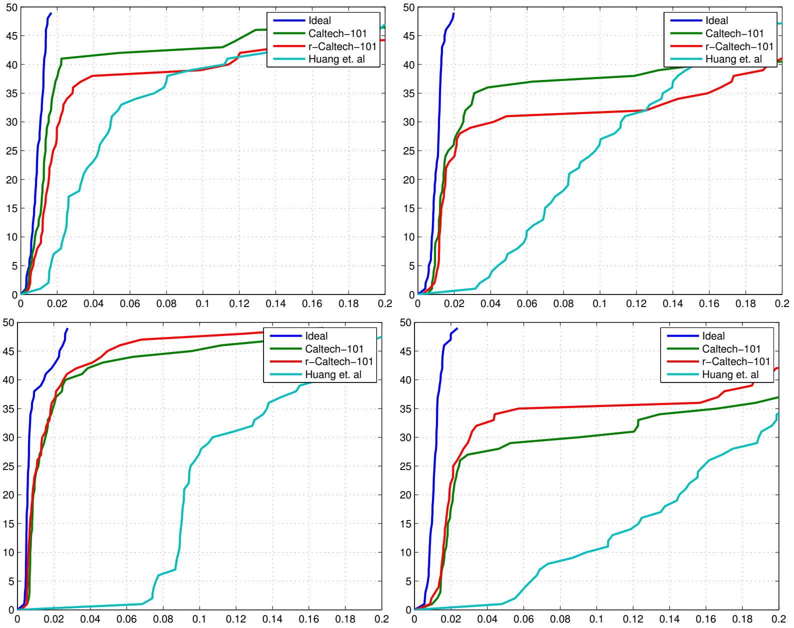

Figure 5: Quantitative results. Cumulative error curves for faces (top left), motorbikes (top right), stop signs (bottom left) and car side (bottom right). Blue: ideal, cyan: Huang *et al*. [13], red: our method for r-Caltech-101, and green: our method for original Caltech-101.

most repeatable local parts.

In addition, it would be intriguing to combine our method with the state-of-the-art congealing methods, such as [5, 13] since a hybrid approach could be useful, for example, in robust and accurate alignment of images in huge existing medical image databases.

References

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

[2] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005.

[3] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial bag-of-features. In *CVPR*, 2010.

[4] Y. Chen, A. Yuille, L. Zhu, and H. Zhang. Unsupervised learning of probabilistic object models (POMs) for object classification, segmentation and recognition. In *CVPR*, 2008.

[5] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for unsupervised alignment of images. In *CVPR*, 2008.

[6] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for large number of images. In *CVPR*, 2009.

[7] G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 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 on Generative-Model Based Vision*, 2004.

[9] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6), 1981.

[10] B. Frey and N. Jojic. Transformed component analysis: Joint estimation of spatial transformations and image components. In *CVPR*, 2000.

[11] B. Frey and N. Jojic. Transformation-invariant clustering and dimensionality reduction using EM. *IEEE PAMI*, 25(1):1–17, 2003.

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2003.

[13] G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[14] T. Jiang, F. Jurie, and C. Schmid. Learning shape prior models for object matching. In *CVPR*, 2009.

[15] T. Kinnunen, J.-K. Kamarainen, L. Lensu, J. Lankinen, and H. Kälviäinen. Making visual object categorization more challenging: Randomized Caltech-101 data set. In *ICPR*, 2010.

[16] I. Kokkinos and A. Yuille. Unsupervised learning of object deformation models. In *ICCV*, 2007.

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

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[19] E.G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE PAMI*, 28(2):236–250, 2006.

[20] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision*, 60(2):91–110, 2004.

[21] M.C.Burl, M.Weber, and P.Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998.

[22] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, 2002.

[23] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 27(10), 2005.

[24] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *CVPR*, 2005.

[25] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int J Comput Vis*, 65(1-2), 2005.

[26] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities of transforms. In *CVPR*, 2000.

[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[28] P. Pritchett and A. Zisserman. Wide baseline stereo. In *ICCV*, 1998.

[29] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE PAMI*, 2008.

[30] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, 2000.

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

[32] A. Vedaldi and S. Soatto. A complexity-distortion approach to joint pattern alignment. In *NIPS*, 2006.

[33] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int J Comput Vis*, 73(2):213–238, 2007.