Distances based on Non-rigid Alignment for Comparison of Different Object Instances

Benjamin Drayer and Thomas Brox

Department of Computer Science,
Centre of Biological Signalling Studies (BIOSS),
University of Freiburg, Germany
{drayer, brox}@cs.uni-freiburg.de

Abstract. Comparison of different object instances is hard due to the large intra-class variability. Part of this variability is due to viewpoint and pose, another due to subcategories and texture. The variability due to mild viewpoint changes, can be normalized out by aligning the samples. In contrast to the classical Procrustes distance, we propose distances based on non-rigid alignment and show that this increases performance in nearest neighbor tasks. We also investigate which matching costs and which optimization techniques are most appropriate in this context.

1 Introduction

A large part of the variability among images of an object class is due to viewpoint and pose. While large differences in viewpoint and pose render the images very different and leave little hope to compare them directly, object samples taken from approximately the same viewpoint share many common features. Descriptors that are invariant to small local deformations, such as HOG, have been the basis for establishing matches between such samples. But is the concept of grids of histograms sufficient?

In this paper, we show that a non-rigid alignment procedure on top of HOG improves the similarity of different object instances from the same class and seen in the same pose; see Fig. 1. This is particularly true if the deformation cost for the alignment is part of the distance.

Alignment procedures have been heavily used in the scope of matching faces. Since (frontal) faces are mostly planar, most face alignment methods focus on rigid or affine alignments, see for instance [7]. For general object classes with more variation, non-rigid alignment is more appropriate, as we show in this paper. Non-rigid alignment has been used also for face alignment [17] but required additional supervision by training fiducial detectors.

Unsupervised non-rigid alignment between object instances, as considered in this paper, is a hard problem, both with regard to the matching cost and with regard to the optimization. On the side of the matching cost, we build upon the idea of whitened HOG features as recently proposed in [6] in the scope of clustering and detection. Moreover, we find that a combination of the l_1 norm and the dot product behaves better than these norms alone.

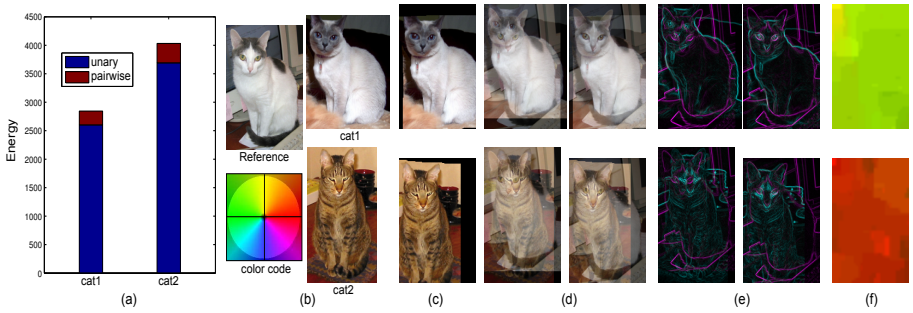

Fig. 1. Illustration and motivation of alignment based distances. (a) Optimum alignment energies. (b) Two cats with different poses to be aligned with the reference cat; color code for the deformation fields. (c) cat1 and cat2 after alignment with the reference. (d) Overlay of the images before and after alignment. (e) Overlay of the gradient images before and after alignment. (f) Estimated deformation fields. Samples that are similar enough to be aligned well yield low energies, whereas samples that cannot be aligned properly yield higher energies.

On the side of optimization, we investigate a set of efficient discrete optimization methods. Matching of different instances has been studied before, e.g., in Berg et al. [1], who solved the corresponding NP-hard optimization problem with a linear programming relaxation. Due to the computational complexity of this approximation, only correspondences of very sparse sets of feature points can be computed. In the context of label transfer between different scenes, SIFT flow [11] has been based on an optimization with belief propagation [15]. Apart from belief propagation, we investigate two other methods: fast primal-dual [8, 9], and α-expansion with non-submodular binary cost functions, so-called fusion moves [10]. We compare the energies obtained with these three techniques, as well as the computation times.

We demonstrate the effect of non-rigid alignment on the distances between instances on four datasets, one on cars, two on cats and one on horses. We aim to find the visually most similar examples relative to a reference image. Each dataset consists of the references and the corresponding ground truth sets of the most similar nearest neighbors. We evaluate various distances with and without alignment. Furthermore, we compare against the scores obtained with an exemplar SVM [12] and with a rigid alignment. The results show that distances based on non-rigid alignment match the annotation much better than distances on raw HOG, whitened HOG features, rigid alignment, or a HOG based exemplar SVM.

2 Non-rigid alignment

For each pair of examples we would like to compute the deformation field that optimally aligns one example with the other. This is much more difficult than,

e.g., optical flow estimation, since there is variation besides the sought deformation, and we do not yet know which features are reliable for matching. We build upon the HOG [3] and the whitened HOG (WHO) descriptor [6]. The WHO descriptor is advantageous as it tends to give more weight to features that are most relevant for the present object class. This feature weighting is potentially also useful for alignment. Indeed, our experiments show an improved performance if the alignment takes into account whitened HOG features.

Unfortunately, the whitening requires inversion of the covariance matrix. Only feature vectors with less than 10000 dimensions can be handled in reasonable time, which corresponds to HOG representations with 16×16 blocks. For finer representations (using more but smaller cells), we must return to the classical HOG representation without whitening. As a consequence, we run a coarse alignment on WHO features and use the resulting deformation field as a soft constraint when optimizing the refined alignment based on HOG. The cost function consists of the matching costs E_D, which aim for maximum feature overlap, and a regularization term E_P that penalizes strong deformations:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + \lambda E_P(\mathbf{u}). \tag{1}$$

This cost function is minimized with respect of the deformation field $\mathbf{u}$. The regularization parameter λ allows to emphasize either the deformation cost or the matching cost. We empirically determined $\lambda = 1.0$ in case of WHO features and $\lambda = 0.2$ for HOG features.

2.1 Matching cost

As matching cost we use a combination of the l_1-norm and the dot product. The advantage of the l_1-norm is its robustness, but at images where we have slightly different features, it can prefer to match a weak feature to the background rather than to the most similar feature; see Fig. 2. On the other hand, the dot product tries to match as many features as possible, but it does not penalize unaligned features. This leads to blurring effects. The combination prefers alignment of the closest features while enforcing one-to-one assignments:

$$E_D(\mathbf{u}) = \sum_{\mathbf{x}} |F_2(\mathbf{x} + \mathbf{u}(\mathbf{x})) - F_1(\mathbf{x})|_1 - \langle F_2(\mathbf{x} + \mathbf{u}(\mathbf{x})), F_1(\mathbf{x}) \rangle \tag{2}$$

where $F(\mathbf{x})$ denotes the feature vector at position $\mathbf{x}$ (the respective cell of the HOG descriptor).

2.2 Deformation cost

For measuring the deformation cost, we use the total variation

$$E_P(\mathbf{u}) = \sum_{\mathbf{x},\mathbf{y} \in \mathcal{N}(\mathbf{x})} |\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{y})|_1 , \tag{3}$$

where $\mathcal{N}(\mathbf{x})$ denotes the 4-connected neighborhood of $\mathbf{x}$. The total variation regularization prefers piecewise constant deformation fields.

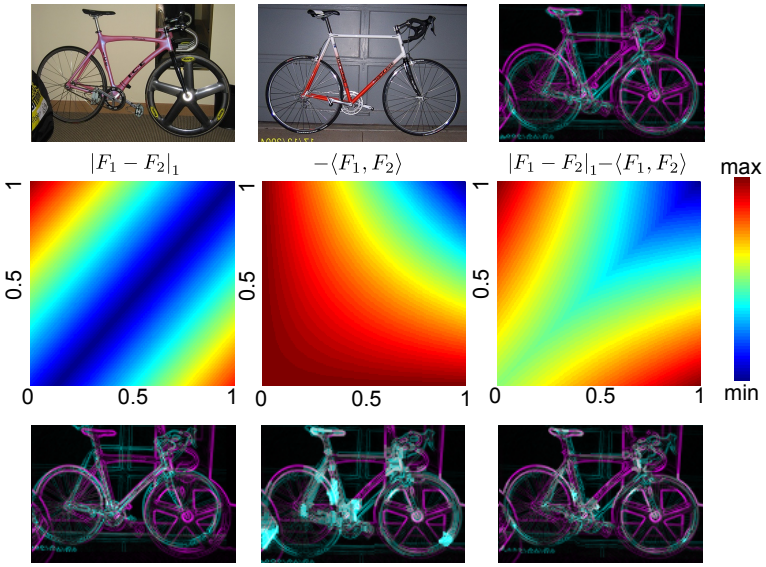

Fig. 2. Top row: Two bikes and the overlay of their gradients. **Center row:** Three different matching costs. **Bottom row:** Resulting alignments. **Left:** With the l1 norm, weak gradients are preferably matched to the background. **Middle:** With the dot product, smearing effects occur because matching to the background does not induce any cost. **Right:** The combination leads to the best alignment.

2.3 Refinement on a finer grid of HOG cells

To exploit both the feature weighting of the WHO features and the higher accuracy of HOG with smaller cells, the initial alignment is obtained by minimizing Eq. 1 based on WHO features. The resulting deformation field $\mathbf{u}_{\text{WHO}}$ serves as a soft constraint in the successive dense alignment on a finer grid of HOG cells. At this fine level, we minimize:

$$E(\mathbf{u}) = \sum_{\mathbf{x}} \beta \, \delta(\mathbf{x}) \, |\mathbf{u}_{\text{WHO}} - \mathbf{u}|_1 + E_D(\mathbf{u}) + \lambda E_P(\mathbf{u})$$
$$\delta(\mathbf{x}) = \begin{cases} 1, & \text{If } \mathbf{u}_{\text{WHO}} \text{ defined at } \mathbf{x} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The function $\delta(\mathbf{x})$ indicates the grid positions where the coarse deformation field $\mathbf{u}_{\text{WHO}}$ is available. The scaling factor $\beta = 0.03$ that regulates the influence of the initial alignment was determined empirically.

2.4 Energy minimization

For the purpose of pairwise comparison, the optimization of the above energies must be fast on one hand, but also sufficiently reliable on the other hand. We

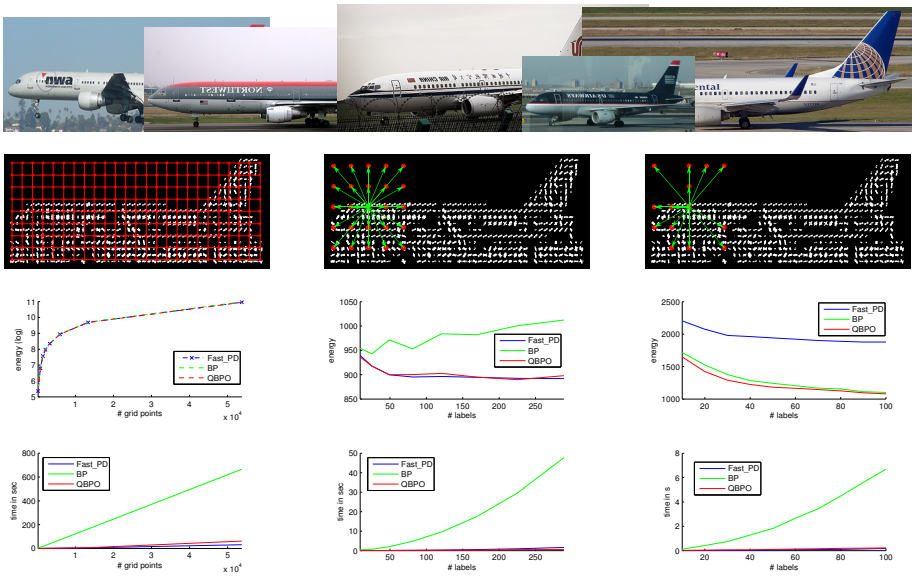

Fig. 3. Top row: Five images, for which we compare the energy minimization techniques. **Second row:** Images in HOG-feature space. In the left column we compare the behavior of the MRF-solvers when changing the number of nodes. In the middle we change the number of labels. The same set of labels is used everywhere, thus the binary subproblems are submodular. In the right column, the label set varies spatially based on the best k displacements. Thus the binary subproblems are no longer submodular. **Third row:** For submodular binary problems, Fast_PD and QPBO perform best. In the non-submodular case, BP and QPBO perform best. **Bottom row:** Run times of the different approaches depending on the number of grid points and labels. Fast_PD and QPBO are much faster than BP.

consider three multi-label MRF solvers: loopy belief propagation, Fast_PD [8, 9] and α-expansion [2] with QPBO [13]. The Fast_PD algorithm solves a sequence of intermediate binary problems with min-cuts. The binary solution is only guaranteed to be optimal, if the binary problem is submodular. Also the α-expansion with QPBO solves a sequence of binary problems, which is done by so-called fusion moves [10]. In contrast to min-cuts, QPBO can solve a larger class of binary problems, the set of pseudo-boolean functions, which includes the submodular problems as a subset. Loopy belief propagation directly optimizes the multi-label problem, but there is neither any guarantee of optimality nor of convergence.

For the experiment in Fig. 3 we aligned five airplane images in HOG space and investigate the behavior of the three approximate optimization techniques, when changing the number of nodes, the number of labels and when we violate the submodularity property. If the labels correspond to the same displacement vectors everywhere, the binary subproblems are all submodular, because assigning the same label induces zero cost, while assigning another label induces higher

cost. In this case, α-expansion with QPBO and Fast_PD minimize the energy equally well. Fast_PD scales a little better with the number of grid points, QPBO scales better with the number of labels. With an adaptive set of displacement vectors, submodularity is lost and Fast_PD does not minimize the energy anymore. Loopy belief propagation is unreliable, due to the missing convergence property, and very slow as the number of grid points or labels increases. The energies in Eq. 1 and 4 use a spatially fixed label set, but due to the better scaling with the number of labels, we used α-expansion with QPBO to compute the non-rigid alignment.

3 Distances based on alignment

A variety of distances can be defined based on the alignment of the previous section. As Fig. 1 indicates, the energy can be used directly as a distance measure. For this direct approach, there are three possibilities that we evaluated: the energy of Eq. 1 using HOG features (E_{HOG}) *or* WHO features (E_{WHO}), and the refinement based approach in Eq. 4 (E_{combi}) that uses both features.

Alternatively, distances can be defined based on the aligned features using the normalized dot product:

$$d(F_1, F_2) = \frac{\langle F_1, F_2 \rangle}{\|F_1\|_2 \cdot \|F_2\|_2}, \tag{5}$$

F_1, F_2 are the HOG or WHO descriptors on the aligned images. The global normalization prevents images with rich gradients to be favored over those with less structure. To this distance we can add the deformation cost λE_P, which provides valuable information on how much the second image needed to be distorted to match the first one. Again, there are three possibilities how to compute the alignment (based on HOG, on WHO, or the combination of both). All these distances are evaluated in the next section.

4 Experiments

4.1 Dataset

We compared various distances between object instances in a simple experiment, where we find for a certain reference image the nearest neighbors according to this distance. To allow for a quantitative evaluation, we considered four datasets, 3D cars from [14], our own cat dataset, Pascal VOC cats [5] and Pascal VOC horses [4]. The 3D cars dataset consists of 10 different cars shown from 8 different viewing angles, and is typically used for viewpoint classification[1]. For each

[1] It is important to note that our experiment is *not* about viewpoint classification. We do not employ a training set to learn the best features to distinguish viewpoints. We are rather interested in an unsupervised definition of distances between examples that resemble human perception and test these distances in a nearest neighbor task.

viewing angle we picked one car as a reference and used the other 9 as ground truth set of nearest neighbors. Our cat dataset consists of 120 cats from Flickr provided by [16], we chose references representing the poses: portrait, walking left, walking front, sitting frontal, sitting left, sitting right and lying right. We manually defined a ground truth set of nearest neighbors for each of these poses. There are some images that do not fit to any of the references. In the same style we added annotation to the two Pascal VOC sets. The 200 cats from Pascal VOC 2006 show a great diversity, consequently we chose the three most frequent patterns (portrait, lying cat, sitting cat) as references together with their sets of nearest neighbors. Among the 724 horses from Pascal VOC 2007, the reference images represent horses from: front, left, right, left-front, right-front, jumping over the fence left and right, begging left and right. On the VOC images, we used the bounding box annotation to clip the image accordingly.

For the evaluation we compute precision and recall, where precision is the percentage of correct nearest neighbors and recall is the percentage of retrieved nearest neighbors. Comparison of samples to the reference based on the evaluated distance yields a ranked list, from which we computed a precision-recall curve. We report the average precision as the area under the precision-recall curve.

4.2 Results

In Table 1, we compare the different distances defined in Section 3. The raw energies do not perform well as they lack global normalization introduced in Eq. 5. On average the alignment helps improve performance for the different features. It works best on cars and horses and does not improve on cats. This is because cats are particularly hard to align due to their large variability, e.g., various textures and large pose variation. The rigidness of the cars makes them the easiest case. Horses also come with non-rigid deformations, but their appearance is not as diverse as that of cats. Fig. 4 shows some qualitative results.

In Table 2 we verified if a simple rigid alignment can achieve a similar performance as a non-rigid alignment. Apart from the alignment model, the definition of the distances is equivalent. For simplicity, we evaluated only the HOG based alignment. The result confirms the need of a non-rigid alignment in case of non-planar objects (unlike faces). Even for cars, distances benefit from a non-rigid alignment.

Moreover, we compared to the score returned by the exemplar SVM (ESVM) [12]. In this approach one reference instance is taken as the only positive example and a linear SVM is trained to separate it from a large set of negative samples. This approach benefits from the SVM figuring out the relevant features that distinguish the positive sample from random samples. We used the reference images of the above datasets as exemplars and used the scores on the other images as similarity measure to pick the nearest neighbors. The result shows that the exemplar SVM is not useful for the purpose of distances. The linear decision function of the SVM lacks expressive power. It is interesting to note that a kernelized version of ESVM with an RBF kernel would be build upon a

Table 1. Comparison of various distances with and without non-rigid alignment in terms of average precision (AP). The distances in the left block use the energies directly, the two blocks in the middle use HOG and WHO features, before and after the alignment. Methods with $+\lambda E_P$ make use of the deformation cost computed during the alignment. In the last two blocks HOG and WHO were both used for alignment, which yields the best results.

	E_{HOG}	E_{WHO}	E_{combi}	HOG	HOG aligned	HOG $+\lambda E_P$	WHOG	WHO aligned	WHO $+\lambda E_P$	HOG combi	WHO combi	HOG combi$+\lambda E_P$	WHO combi $+\lambda E_P$
Cars	30.39	30.4	31.74	30.79	44.94	43.57	28.09	30.05	30.45	39.42	30.05	**46.01**	33.8
Cats own	13.6	13.78	13.81	31.18	31.97	32.23	33.04	30.66	30.99	32.93	30.66	**33.41**	32.29
Cats Pascal	6.55	6.61	6.56	33.16	31.81	31.05	31.32	30.97	31.24	27.82	27.42	32.58	**33.17**
Horse Pascal	4.32	4.31	4.22	29.49	36.47	37.38	33.34	35.43	36.42	36.87	35.43	**38.83**	33.8
Mean	13.72	13.78	14.08	31.16	36.3	36.1	31.45	31.78	32.28	34.26	30.89	**37.71**	33.27

Table 2. Performance of ESVM [12], the rigid alignment and non-rigid aligned HOG-features. The non-rigid alignment consistently shows better AP.

	Cars	Cats own	Cats Pascal	Horse Pascal	Mean
ESVM [12]	24.07	16.83	10.68	19.08	17.67
rigid alignment	41.25	27.76	27.05	29.77	31.46
HOG aligned	44.94	31.97	31.81	36.47	**36.3**

distance between the reference and negative samples, which takes us back to the definition of appropriate distances.

5 Conclusions

We have suggested distances between different object instances based on non-rigid alignment. We showed in a nearest neighbor experiment that distances based on non-rigid alignment perform better than distances based on an rigid alignment or no alignment at all. Moreover, thanks to an efficient optimization, non-rigid alignments can be computed also on larger datasets in reasonable time. Pairwise distances appear in many learning problems, such as clustering or kernel based classifiers. Hence, we believe that alignment based distances can have a positive effect in several applications.

Acknowledgements

This study was supported by the Excellence Initiative of the German Federal and State Governments (EXC 294) and by the ERC Starting Grant VIDEOLEARN.

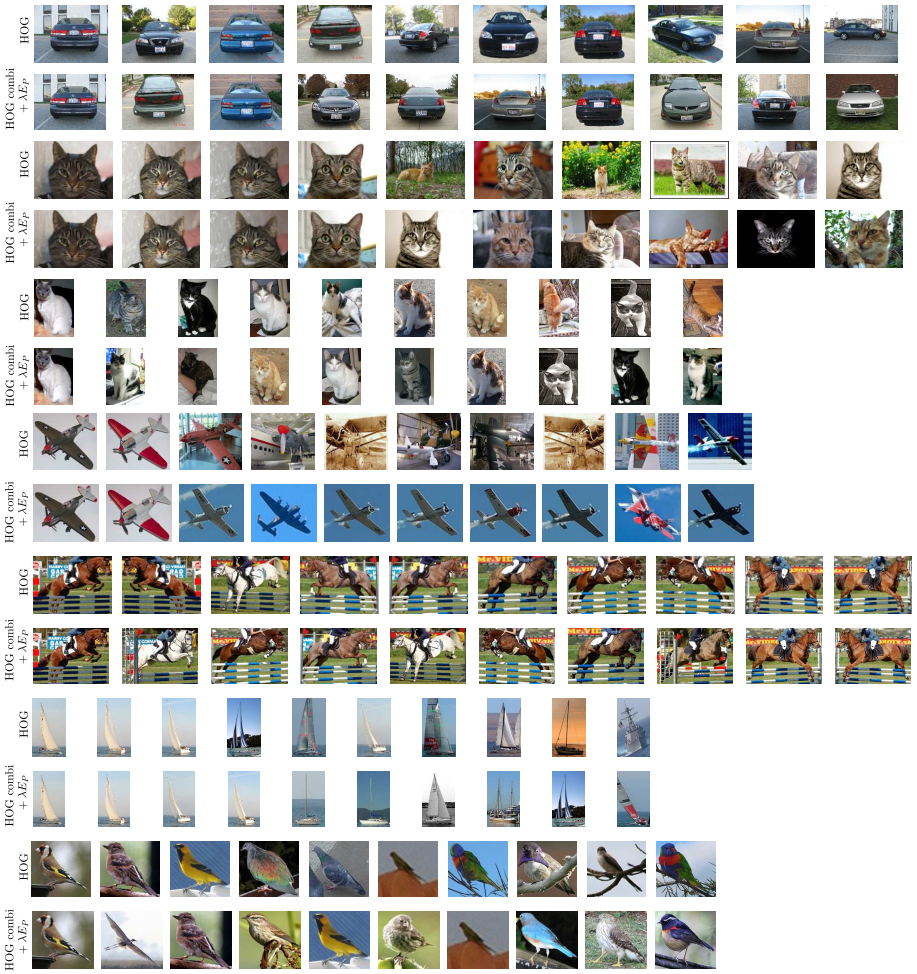

Fig. 4. Qualitative comparison between distance measures on non-aligned (HOG) and aligned (HOG combi+λE_P) images. The left column shows the reference image, the most similar images are ordered from left to right. For each reference image, we show the 9 most similar images with respect to non-aligned HOG features and aligned HOG features. The first example is from the car database [14], the second is from our own cat dataset, the third shows cats from Pascal VOC 2006 [5]. The remaining examples are from Pascal VOC 2007 [4]. In general, the samples found with the alignment based distance are more meaningful.

References

1. Berg, A.C., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: In CVPR. pp. 26–33 (2005)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1222–1239 (Nov 2001), http://dx.doi.org/10.1109/34.969114
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: International Conference on Computer Vision & Pattern Recognition. vol. 2, pp. 886–893 (June 2005)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
5. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
6. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: European Conference on Computer Vision (ECCV) (2012)
7. Huang, G.B., Mattar, M.A., Lee, H., Learned-Miller, E.G.: Learning to align from scratch. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS (2012)
8. Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. IEEE Trans. Pattern Anal. Mach. Intell. 29(8), 1436–1453 (Aug 2007)
9. Komodakis, N., Tziritas, G., Paragios, N.: Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. Comput. Vis. Image Underst. 112(1), 14–29 (Oct 2008), http://dx.doi.org/10.1016/j.cviu.2008.06.007
10. Lempitsky, V.S., Rother, C., Roth, S., Blake, A.: Fusion moves for markov random field optimization. IEEE Trans. Pattern Anal. Mach. Intell. 32(8), 1392–1405 (2010)
11. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: Dense correspondence across different scenes pp. 28–42 (2008)
12. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
13. Rother, C., Kolmogorov, V., Lempitsky, V., Szummer, M.: Optimizing binary mrfs via extended roof duality. Tech. rep., In Proc. CVPR (2007)
14. Savarese, S., Fei-Fei, L.: 3d generic object categorization, localization and pose estimation. In: IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil (October 2007)
15. Shekhovtsov, A., Kovtun, I., Hlaváč, V.: Efficient mrf deformation model for non-rigid image matching. In: In IEEE Transactions on International Conference on Pattern Recognition (2007)
16. Zhang, W., Sun, J., Tang, X.: Cat head detection - how to effectively exploit shape and texture features. In: In ECCV (2008)
17. Zhu, J., Gool, L.J.V., Hoi, S.C.H.: Unsupervised face alignment by robust nonrigid mapping. In: ICCV. pp. 1265–1272. IEEE (2009)