Dense Semantic Correspondence where Every Pixel is a Classifier

Hilton Bristow,[1] Jack Valmadre[1] and Simon Lucey[2]
[1]Queensland University of Technology, Australia
[2]Carnegie Mellon University, USA

Abstract

Determining dense semantic correspondences across objects and scenes is a difficult problem that underpins many higher-level computer vision algorithms. Unlike canonical dense correspondence problems which consider images that are spatially or temporally adjacent, semantic correspondence is characterized by images that share similar high-level structures whose exact appearance and geometry may differ.

Motivated by object recognition literature and recent work on rapidly estimating linear classifiers, we treat semantic correspondence as a constrained detection problem, where an exemplar LDA classifier is learned for each pixel. LDA classifiers have two distinct benefits: (i) they exhibit higher average precision than similarity metrics typically used in correspondence problems, and (ii) unlike exemplar SVM, can output globally interpretable posterior probabilities without calibration, whilst also being significantly faster to train.

We pose the correspondence problem as a graphical model, where the unary potentials are computed via convolution with the set of exemplar classifiers, and the joint potentials enforce smoothly varying correspondence assignment.

1. Introduction

Unlike canonical dense correspondence problems which consider images that are spatially (stereo) or temporally (optical flow) adjacent, semantic correspondence is characterized by images that stem from the same visual class (*e.g.* elephants, lammergeiers, carlined streets) whilst exhibiting individual appearance and geometric properties.

For example, given two images of elephants (see Figure 1), we would like to predict where each pixel on the first elephant corresponds to on the second. This is particularly challenging because the space of elephants exhibits significant intra-class appearance and geometric variation. A related problem is that of pose estimation [17, 24], which considers a smaller fixed set of landmarks stemming from a labelled dataset of a known object class. From this dataset, one can learn (i) the geometric dependency between landmarks, and (ii) local detectors that discriminate the appearance of each landmark from the background. When presented with a new image, one can then estimate the landmark locations by solving a graphical inference problem.

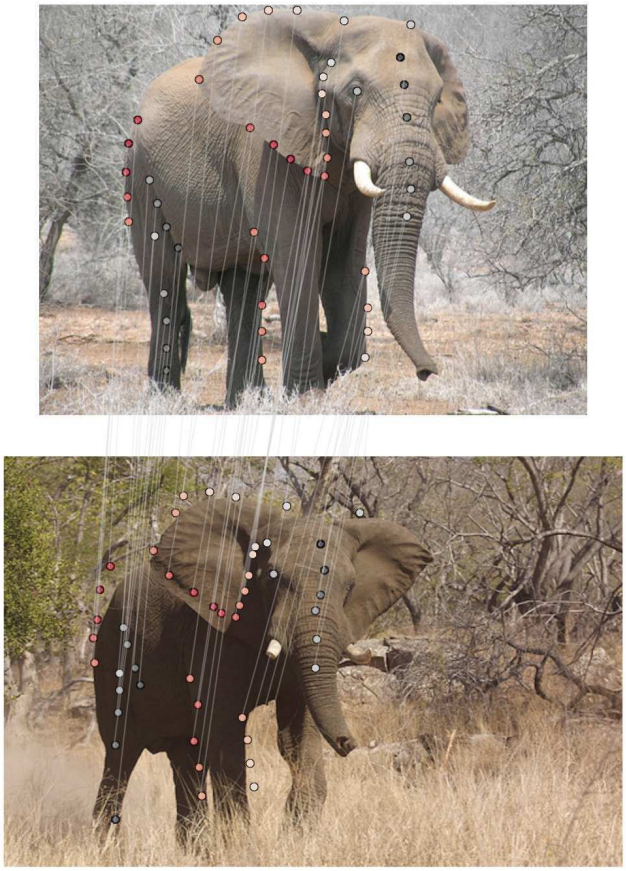

Figure 1. Dense semantic correspondence estimates how points are related between images that stem from the same visual class. Here, we wish to predict where each pixel on the first elephant corresponds to on the second, whilst being robust to appearance, pose and background variation. The points labelled are representative of the dense correspondence field estimated by our method.

Liu *et al.*'s seminal work of SIFT Flow [12] established that a similar strategy could be applied for estimating dense semantic correspondence between two images stemming from the same semantic class. There are three complicating factors however: (i) learning geometric dependencies between landmarks is impossible from only a single example, (ii) learning local detectors is problematic due to the lack of positive training samples, and (iii) computational complexity is a major concern as we are treating each pixel coordinate within the image as a landmark. Liu *et al.* proposed to circumvent these problems by assuming the dense geometric dependencies in an image can be adequately governed by a variational regularizer, and that accurate local detections between semantically similar images can be attained through the L_1 distance between SIFT descriptors. Since there is no learning required, this can be performed in a computationally tractable manner.

In this paper, we explore the possibility of actually learning a discriminative detector at every pixel coordinate in an image. Motivated by object detection literature, we learn a linear classifier per pixel in the reference image and apply it in a sliding-window manner to the target image to produce a match likelihood estimate. Learning a multitude of linear detectors such as exemplar support vector machines (SVMs) has typically had two issues: (i) each detector must parse the negative set, often with hard-negative mining techniques, leading to long training times, which makes training a classifier for every pixel in an image intractable, and (ii) since the scale of the outputs depends on the margin, the output confidences of two different SVMs are not directly comparable.

We leverage recent work on learning detectors quickly with linear discriminant analysis (LDA), by collecting negative statistics across a large number of images in a pre-training phase. Learning a new exemplar detector then involves a single matrix-vector multiplication. Since LDA uses a generative model of the class distributions, the posterior probabilities provide a quantity that is comparable between detectors. This allows us to estimate both the likelihood of matches for each pixel individually, and also a global belief of match quality.

2. Prior Art

Canonical correspondence problems such as stereo and optical flow typically rely on simple (dis-)similarity metrics to describe the likelihood of two pixels matching. In the original work of Horn and Schunck [7], this was Euclidean distance on raw pixel intensities, which manifested a brightness constancy assumption.

Since then, significant literature has focused on determining robust metrics under increasingly adverse conditions - from non-rigid deformations and occlusions, to non-global intensity, constrast and colorimetric changes [1, 14, 19, 20]. Importantly, however, all of these works assume the images being observed stem from the same underlying scene.

SIFT Flow first introduced the idea of semantic correspondence *across* scenes [12]. While the method uses a simple L_1 metric, the images are represented in dense SIFT space typically associated with sparse keypoint matching.[1] This sacrifices some localization accuracy for improved geometric invariance. We maintain, however, that similarity metrics are insufficient for estimating the likelihood of pixels matching between different scenes. Instead, we advocate the use of classifiers, as per deformable face fitting and pose estimation literature, except where a classifier is trained *per pixel*.

We leverage recent work on rapid estimation of LDA classifiers [4, 22] to achieve this goal, though fast correlation filter estimation [5] is potentially equally applicable. The method we present is largely agnostic to the objective used to learn the linear detectors (*e.g.* SVM, LDA, correlation filters), however LDA classifiers are attractive in producing globally interpretable outputs across pixels, and requiring only a single matrix-vector multiplication to train, which is critical to learning $> 10,000$ classifiers per image.

A number of dense correspondence methods have made use of discriminative pre-training [11, 18, 20], with the recent work of [10] being particularly relevant to our discussion. In this work, a classifier of the form $f(\Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2))$ is trained to predict a (binary) likelihood of two pixels matching. Intuitively, the classifier learns the modes and scale of variation in the underlying feature space Φ that are important and those that are distractors. Training is fully supervised from groundtruth optical flow data.

Like SIFT Flow, [10] formulate the correspondence objective as a graphical model ([8, 9] respectively). This has the distinct advantage over variational methods of permitting very large displacements and arbitrarily complex data terms, at the expense of requiring simple regularizers to keep inference tractable. More recently, a number of variational methods have used sparse descriptor matching to anchor larger displacements [2, 23]. While both methods use robust SIFT descriptors for keypoint matching, in a semantic correspondence setting the best match is infrequently the true correspondence, leading to poor initialization of the densification stage.

[1] Feature representation and similarity metric are intrinsically related, since $f(\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2)) = g(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

Mobahi *et al*. [15] present a parametric approach to the general alignment problem, in a similar vein to congealing [3] or RASL [16]. Parametric methods generally have limited degrees of freedom, which enables strong coarse localization, at the expense of fine-grained localization accuracy.

Finally, the recent work of FlowWeb [25], is particularly complementary to the ideas presented in this paper, since they are agnostic to the method of appearance matching, and we are agnostic to the method of regularization or global matching consistency. We use the variational regularizer of SIFT Flow in our experiments, though the ideas we present are equally applicable to FlowWeb.

3. Dense Semantic Correspondence

Given two images, $\mathcal{I}_A \in \mathcal{R}^{MN}$ and $\mathcal{I}_B \in \mathcal{R}^{PQ}$, and a discrete set of points $\boldsymbol{x} = [x_1, x_2, \ldots, x_{MN}]$, dense semantic correspondence involves minimizing the inverse fitting problem,

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{MN} f_i(\boldsymbol{x}_i) + \lambda g(\boldsymbol{x}) \tag{1}$$

where f is the unary function that evaluates the likelihood of a particular assignment for each $\boldsymbol{x}_i$ based on the image content, and g is a regularizer which enforces constraints on the joint configuration of the points. In semantic correspondence, the unary function must be a good indicator of semantic similarity, and so must be robust to significant intra-class variation. In the framework we adopt, there are no constraints on its complexity or properties.

SIFT Flow [12] adopts a unary of the form,

$$f_i(\boldsymbol{x}_i) = h(i, \boldsymbol{x}_i) = ||\Phi_A(i) - \Phi_B(\boldsymbol{x}_i)||_1 \tag{2}$$

where $\Phi_A(\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i; \mathcal{I}_A)$ is a feature representation of the image $\mathcal{I}_A$ evaluated at the point $\boldsymbol{x}_i$. Φ subsumes the exact detail of the feature extraction, which may produce vector-valued output. For our LDA classifiers, we extract features from a window of pixels around $\boldsymbol{x}_i$ and return a multi-channel patch centered at $\boldsymbol{x}_i$.

In [10], the L_1 norm on the difference between features is replaced with a more general learned representation,

$$h(i, \boldsymbol{x}_i) = H(\Phi_A(i) - \Phi_B(\boldsymbol{x}_i)) \tag{3}$$

In both formulations, however, the unary function is a stationary kernel. This implies a feature space capable of producing similar outputs for semantically similar inputs. Finding such a feature embedding is a difficult task in general, and as a result significant object detection literature has focussed on learning classifiers to distinguish classes instead.

The use of classifiers has two distinct advantages over stationary kernels for describing match likelihood. First, linear classifiers define half-spaces in which samples are either classified as positive or negative. Thus two points with dissimilar appearances can still be afforded a high match likelihood. Second, the importance of different dimensions in the feature space can be learned from data.

In this paper, we advocate a unary function of the form,

$$f_i(\boldsymbol{x}_i) = h(i, \boldsymbol{x}_i) = \mathbf{w}_A(i)^T \Phi_B(\boldsymbol{x}_i) \tag{4}$$

where $\mathbf{w}_A(i)$ is a linear classifier trained to predict correspondences to pixel i in $\mathcal{I}_A$, with ideal response,

$$\mathbf{w}_A(i)^T \Phi_B(\boldsymbol{x}_i) = \begin{cases} 1 & \boldsymbol{x}_i = \boldsymbol{x}_i^* \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

This is traditional binary classification, where the positive class contains the reference pixel, and its true correspondence in the target image, and the negative class contains all other pixels. Since the correspondence in the target image is not known *a priori* however, we rely on the classifier $\mathbf{w}_A(i)$ to generalize from a single training example: $\Phi_A(i)$. This is known as exemplar-based classification [13].

The challenge is how to rapidly estimate thousands of exemplar classifiers per image in reasonable time. The remainder of this section focuses on addressing that challenge, and a number of interesting properties that arise from our approach.

3.1. Learning Detectors Rapidly using Structured Covariance Matrices

Linear classifiers have a rich history in computer vision, not least because of their interpretation and efficient implementation as a convolution operation. Support vector machines have proven particularly popular, due to their elegant theoretical interpretation, and impressive real-world performance, especially on object and part detection tasks. A challenge for any object detection problem is how to treat the potentially infinite negative set (comprising all incorrect correspondences in our case). Object detection methods using support vector machines employ hard negative mining strategies to search the negative set for difficult examples, which can be represented parametrically in terms of the decision hyperplane. This feature is also their limitation for rapid estimation of many classifiers, since each classifier must reparse the negative set looking for

hard examples – knowing one classifier does not help in estimating another.[2]

Linear Discriminant Analysis (LDA), on the other hand, summarizes the negative set into its mean and covariance. The parameters $\mathbf{w}$ of the decision hyperplane $\mathbf{w}^T\boldsymbol{x} = c$ are learned by solving the system of equations,

$$\mathbf{S}\mathbf{w} = \mathbf{b} \tag{6}$$

where $\mathbf{S}$ is the joint covariance of both classes and $\mathbf{b} = \bar{\boldsymbol{x}}_{\text{pos}} - \bar{\boldsymbol{x}}_{\text{neg}}$ is the difference between class means. [4] made two key observations about LDA: (i) if the number of positive examples is small compared to the number of negative examples, the joint covariance $\mathbf{S}$ can be approximated by the covariance of the negative distribution alone, and reused for all positive classes, and (ii) gathering and storing the covariance can be performed efficiently if the negative class is shift invariant (*i.e.* a translated negative example is still a negative example).

This second fact implies stationarity of the negative distribution, where the covariance of two pixels is defined entirely by their relative displacement. Importantly, both [4] and [6] showed that the performance of linear detectors learned by exploiting the stationarity of the negative set is comparable to SVM training with hard negative mining.

The covariance $\mathbf{S}$ can be constructed from a relative displacement tensor, according to,

$$\mathbf{S}_{(u,v,p),(i,j,q)} = g[i-u, j-v, p, q] \tag{7}$$

where i, j, u, v index spatial co-ordinates, and p, q index channels. We call the maximum displacement observed $\text{abs}(i-u)$, $\text{abs}(j-v)$ the bandwidth of the tensor. Also note that stationarity only exists spatially – cross-channel correlations are stored explicitly. The storage of g thus scales quadratically in both bandwidth and channels, though since the detectors we consider are typically small-support, we can entertain feature representations with large numbers of channels (*i.e.* SIFT).

In order to compute $\mathbf{S}$, and thus g, we gather statistics across a random subset of $50,000$ images from ImageNet. We precompute the covariance matrix of the chosen detector size (typically 5×5) and factor it with either a Cholesky decomposition, or its explicit inverse. The constructed covariance $\mathbf{S}$ is a sample covariance matrix estimated from missing data, so whilst it is symmetric by construction, it is not strictly positive-semidefinite. To ensure positive-definiteness, we subtract the minimum of zero and the minimum eigenvalue from the diagonal, *i.e.* $(\mathbf{S} - \min(0, \lambda_{\min}) \cdot \mathbf{I})^{-1}$.

For each pixel in the reference image, we compute,

$$\mathbf{w}_A(i) = \mathbf{S}^{-1}(\bar{\boldsymbol{x}}_{\text{pos}} - \bar{\boldsymbol{x}}_{\text{neg}}) \tag{8}$$

which involves a single vector substraction and matrix-vector multiplication, where,

$$\bar{\boldsymbol{x}}_{\text{pos}} = \Phi_A(\boldsymbol{x}_i) \tag{9}$$

The likelihood estimate for the i-th reference point across the target image can be performed via convolution over the discretize pixel grid,

$$f_i(\boldsymbol{x}) = \mathbf{w}_A(i) * \Phi_B(\boldsymbol{x}) \tag{10}$$

Since storing the full unary is quadratic in the number of image pixels (quartic in the dimension), we perform coarse-to-fine or windowed search as per SIFT Flow [12].

3.2. Posterior Probability Estimation

Linear Discriminant Analysis (LDA) has the attractive property of generatively modelling classes as Gaussian distributions with equal (co-)variance. This permits direct computation of posterior probabilities via application of Bayes' Rule:

$$p(C_{\text{pos}}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_{\text{pos}})\, p(C_{\text{pos}})}{\sum_{n\in\{\text{pos,neg}\}} p(\boldsymbol{x}|C_n)\, p(C_n)} \tag{11}$$

where,

$$p(\boldsymbol{x}|C_n) = \frac{1}{(2\pi)|\mathbf{S}|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_n)^T\mathbf{S}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_n)} \tag{12}$$

With some manipulation, the posterior of Equation (11) can be expressed as,

$$p(C_{\text{pos}}|\boldsymbol{x}) = \frac{1}{1+\mathrm{e}^{-\mathbf{y}}} \tag{13}$$

$$\mathbf{y} = \boldsymbol{x}^T\mathbf{S}^{-1}(\bar{\boldsymbol{x}}_{\text{pos}} - \bar{\boldsymbol{x}}_{\text{neg}}) \tag{14}$$

$$+ \tfrac{1}{2}\bar{\boldsymbol{x}}_{\text{pos}}^T\mathbf{S}^{-1}\bar{\boldsymbol{x}}_{\text{pos}} - \tfrac{1}{2}\bar{\boldsymbol{x}}_{\text{neg}}^T\mathbf{S}^{-1}\bar{\boldsymbol{x}}_{\text{neg}} \tag{15}$$

$$+ \ln\left(\frac{p(C_{\text{pos}})}{p(C_{\text{neg}})}\right) \tag{16}$$

Equation (13) takes the form of a logistic function, which maps the domain $(-\infty, \infty)$ to the range $(0 \ldots 1)$.

The logistic function is typically used to convert SVM outputs to probabilistic estimates, however a "calibration" phase is required to learn the bias and

[2] This is not strictly true. Warm starting an SVM from a previous solution, especially in exemplar SVMs where only a single positive example changes, can induce a significant empirical speedup, however is unlikely to change the $O()$ complexity of the algorithm.

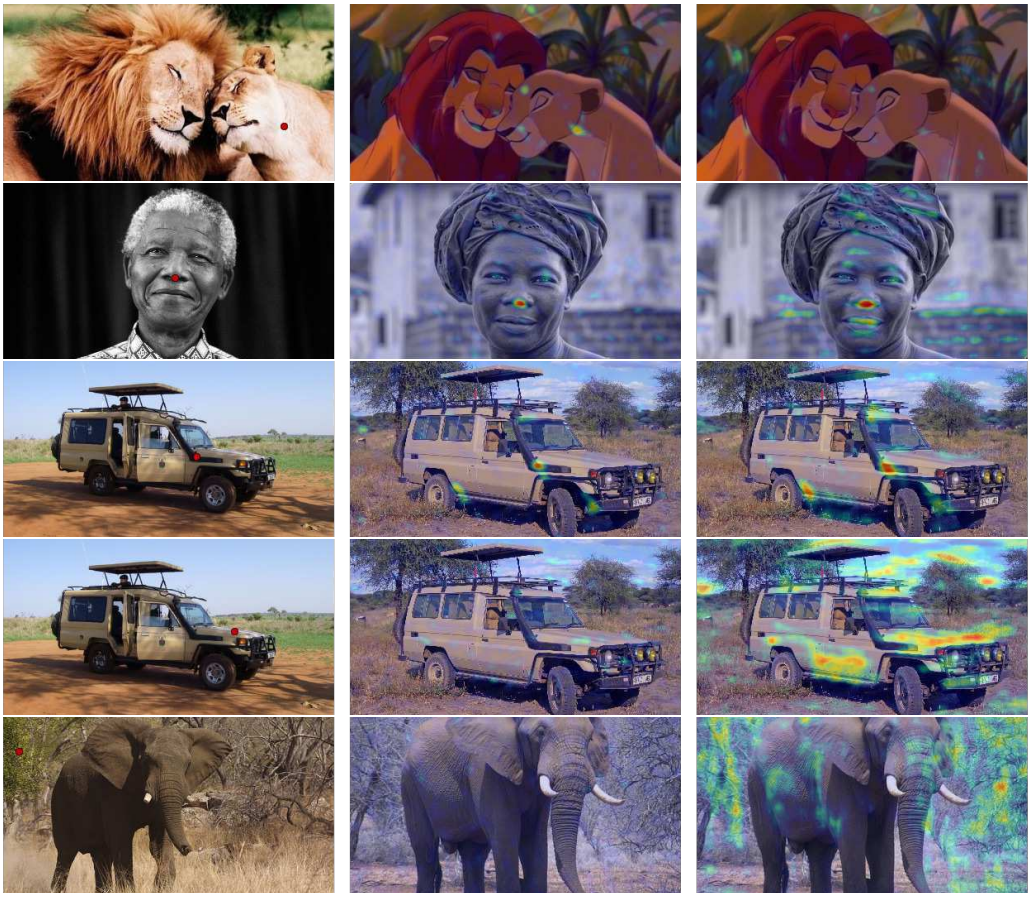

Figure 2. From left to right: (a) reference image with reference point labelled in red, and posterior estimates for (b) LDA and (c) L_1 norm. We present a range of points, from distinctive to indistinctive or background. LDA and L_1 norm have similar likelihood quality for distinctive points, but LDA consistently offers better rejection of incorrect matches and background content.

variance of each SVM in the ensemble so their outputs are comparable. With LDA, these parameters are derived directly from the underlying distributions.

Equation (14) is the canonical response to the LDA classifier, Equation (15) represents the bias of the distributions, and Equation (16) is the ratio of prior probabilities of the classes. This must be determined by cross-validation (once, not for each classifier), based on the desired sensitivity to true versus false positives.

By completing the squares in Equation (15), we yield the final expression for computing the posterior probability,

$$\begin{aligned}\mathbf{y} &= (\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{x}}_{\text{pos}} + \bar{\boldsymbol{x}}_{\text{neg}}))^T \mathbf{S}^{-1}(\bar{\boldsymbol{x}}_{\text{pos}} - \bar{\boldsymbol{x}}_{\text{neg}}) + \mu \\ &= (\boldsymbol{x} - \tfrac{1}{2}(\bar{\boldsymbol{x}}_{\text{pos}} + \bar{\boldsymbol{x}}_{\text{neg}}))^T \mathbf{w}_i + \mu \qquad (17)\end{aligned}$$

The implication of Equation (17) is that it is no more expensive to compute probability estimates than to just evaluate the classifier – the computation is still dominated by the single matrix-vector product required to learn the classifier.

Figure 2 illustrates a representative set of likelihood estimates output by our method and SIFT Flow respectively. LDA typically has tighter responses around the true correspondence, and better suppression of false positives, especially on background content that has no clear correspondence.

4. Evaluation

In order to evaluate the efficacy of our method, we first wanted to understand how well human annotators perform at semantic labelling tasks. Since we are primarily interested in estimating correspondences for reconstruction-type objectives, we gathered 20 pairs of images from visual object categories which exhibit anatomical correspondence, including an assortment of animals, trucks, faces and people. Given a set of sparsely selected keypoints in the first image of each pair, 8 human annotators were tasked with labelling the corresponding points in the second image. A representative subset of the data is shown in Figure 3.

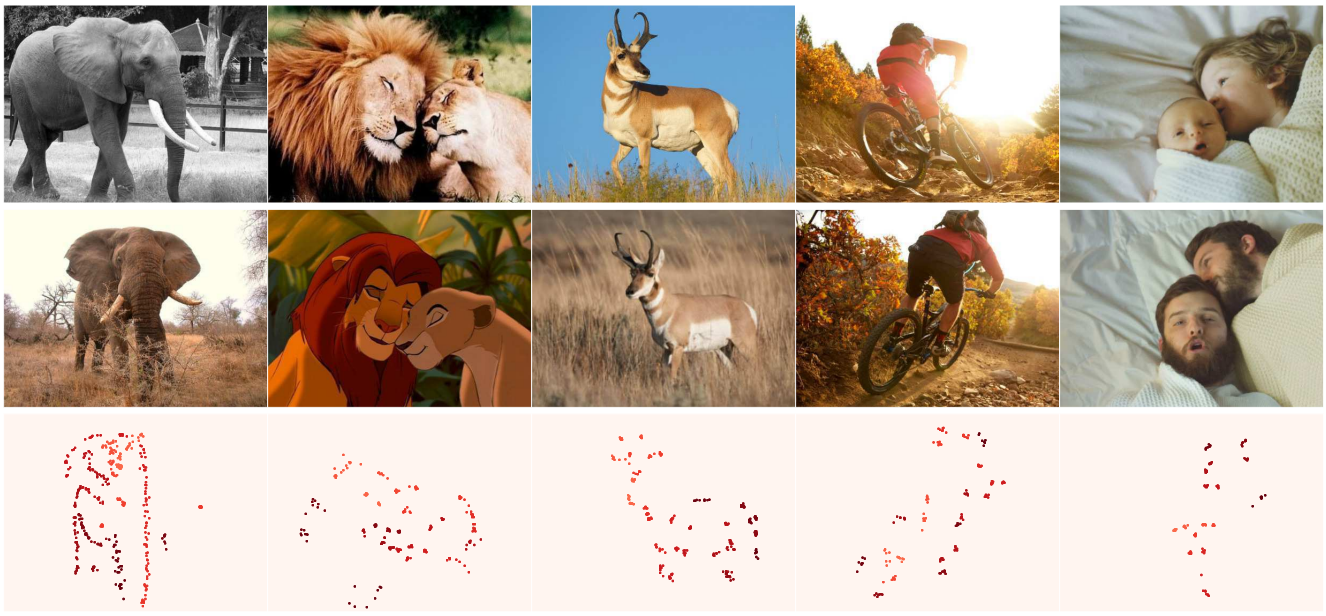

Figure 3. A representative subset of the groundtruth dataset. From top to bottom: (a) the source images, (b) the target images, and (c) the distribution of points selected by the human annotators on the target images. The structure of the object is often clearly discernible from the annotations alone.

A similar experiment was performed in [12], however they focussed on correspondences across *scenes*, which often have no clear correspondence, even for human annotators. In contrast, the agreement on our dataset is high, with a natural increase in uncertainty from corner features, to edges and textureless regions.

In recognizing that not all features are equally distinctive, we measure distance from estimated points $\boldsymbol{x}_i$ to the groundtruth using Mahalanobis distance,

$$d_i(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \mu_i)^T \mathbf{S}_i^{-1} (\boldsymbol{x}_i - \mu_i)} \qquad (18)$$

where μ_i and $\mathbf{S}_i$ are the $2D$ mean and covariance of the groundtruth labellings across annotators. [21] motivate a similar procedure for human pose estimation. This metric has two advantages over Euclidean distance: (i) it takes into account spatial and directional uncertainty (*e.g.* correspondences are afforded some slack along an edge, but not perpendicular to it), and (ii) it is resolution independent, since distance is measured in standard deviations.

Our dataset and metric therefore sets a higher standard for what is considered a good correspondence, both empirically and qualitatively (since readers can accurately discriminate good from poor results). All results presented in the following section are measured under this metric.

4.1. Experiments

In all of our experiments we resize the source (A) and target (B) image so $\max(M, N) = 150$, preserving the aspect ratio, and extract densely sampled SIFT features.

The stationary distribution (mean and covariance) of SIFT features is estimated from $50,000$ randomly sampled images from ImageNet. Classifiers with spatial support 1×1, 3×3, 5×5, 7×7 and 9×9 were evaluated. The different sizes tradeoff speed, localization accuracy and generalization. We found 5×5 classifiers provided a good balance between these tradeoffs, and the results throughout our paper use this support.

While the LDA likelihoods are more computationally demanding to compute than L_1-norm likelihoods, the construction and application of the classifiers can be accelerated with BLAS. Estimating $10,000$ 5×5 classifiers and applying them in a sliding window fashion to a 80×125 SIFT image (with 128 channels) takes approximately 6 seconds.

We apply our LDA-based correspondence method in the same graphical model framework as SIFT Flow. We use a coarse-to-fine scheme to handle inference over larger images, and grid searched the hyperparameters for both LDA and L_1 based unary functions. Results are shown in Figure 4.

We display the cumulative density for increasing number of standard deviations from groundtruth (*i.e.* fraction of points falling within an increasing radius

from groundtruth). As a baseline, we simply set $\boldsymbol{x}_i = i$,[3] which acts as a proxy to the global alignment bias of the dataset (small flow assumption). In addition to SIFT Flow, we also compare our method to a leading optical flow method, Deep Flow [23].

We truncate the CDF due to the long tails for all methods compared. This is an artefact of the non-global regularization schemes, which allow some points to be arbitrarily far from groundtruth without affecting others. Finally, in Figure 5 we illustrate a number of exemplar correspondences to show the visual quality of matches produced by our method.

5. Conclusion

In this paper we motivated the application of dense semantic correspondence for a range of computer vision problems which currently rely on synthetic data or specialized imaging devices. In contrast to existing correspondence methods, which typically use similarity kernels, we proposed using exemplar classifiers for describing the likelihood of two points matching. We showed that LDA classifiers exhibit 3 desirable properties: (i) higher average precision than simple measures of image similarity such as the L_1 norm, (ii) significantly faster training than exemplar SVMs, and (iii) estimates of match confidence that are directly comparable across pixels.

We presented a small semantic correspondence dataset and metric in a bid to measure the performance of different methods in a quantifiable manner, and showed that under this metric our classifier-based approach offered improvements over the L_1 norm, within the same SIFT Flow optimization framework. The qualitative results illustrate our method's ability to estimate high-quality dense semantic correspondences.

References

[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. *European Conference on Computer Vision (ECCV)*, 2004. 2

[2] T. Brox, J. Malik, and C. Bregler. Large displacement optical flow. *International Conference of Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[3] M. Cox, S. Sridharan, and S. Lucey. Least-squares congealing for large numbers of images. *International Conference on Computer Vision (ICCV)*, 2009. 3

[4] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. *European Conference on Computer Vision (ECCV)*, 2012. 2, 4

[5] J. a. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-Speed Tracking with Kernelized Correlation Filters. *Pattern Analysis and Machine Intelligence (PAMI)*, 2014. 2

[6] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista. Beyond Hard Negative Mining: Efficient Detector Learning via Block-Circulant Decomposition. *International Conference on Computer Vision (ICCV)*, 2013. 4

[7] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 1981. 2

[8] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence (PAMI)*, 2006. 2

[9] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2011. 2

[10] L. Ladicky, C. Häne, and M. Pollefeys. Learning the Matching Function. *arXiv preprint*, 2015. 2, 3

[11] Y. Li and D. Huttenlocher. Learning for optical flow using stochastic optimization. *European Conference on Computer Vision (ECCV)*, 2008. 2

[12] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *Pattern Analysis and Machine Intelligence (PAMI)*, 2011. 2, 3, 4, 6, 8

[13] T. Malisiewicz, A. Gupta, and A. a. Efros. Ensemble of exemplar-svms for object detection and beyond. *International Conference on Computer Vision (ICCV)*, 2011. 3

[14] Y. Mileva, A. Bruhn, and J. Weickert. Illumination-Robust Variational Optical Flow with Photometric Invariants. *Pattern Recognition*, 2007. 2

[15] H. Mobahi and W. T. Freeman. A Compositional Model for Low-Dimensional Image Set Representation. *International Conference of Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[16] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *Pattern Analysis and Machine Intelligence (PAMI)*, 2012. 3

[17] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose Machines: Articulated Pose Estimation via Inference Machines. *European Conference on Computer Vision (ECCV)*, 2014. 1

[18] S. Roth and M. Black. On the spatial statistics of optical flow. *International Conference on Computer Vision (ICCV)*, 2005. 2

[19] S. M. Seitz and S. Baker. Filter flow. *International Conference on Computer Vision (ICCV)*, 2009. 2

[20] D. Sun, S. Roth, J. P. Lewis, and M. J. Black. Learning optical flow. *European Conference on Computer Vision (ECCV)*, 2008. 2

[21] J. Tompson, R. Goroshin, A. Jain, Y. Lecun, and C. Bregler. Efficient Object Localization Using Convolutional Networks. *arXiv preprint*, 2015. 6

[22] J. Valmadre, S. Sridharan, and S. Lucey. Learning detectors quickly using structured covariance matrices. In *Asian Conference on Computer Vision (ACCV)*, 2014. 2

[23] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large Displacement Optical Flow with Deep Matching. *International Conference on Computer Vision (ICCV)*, 2013. 2, 7, 8

[24] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[25] T. Zhou, Y. J. Lee, S. X. Yu, U. C. B. Icsi, and A. A. Efros. FlowWeb: Joint Image Set Alignment by Weaving Consistent, Pixel-wise Correspondences. *International Conference of Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[3] For images of different sizes, we set $\boldsymbol{x}_i = \mathcal{W}(i)$ where $\mathcal{W}$ is a function that maps the span of $\mathcal{I}_A$ to $\mathcal{I}_B$.

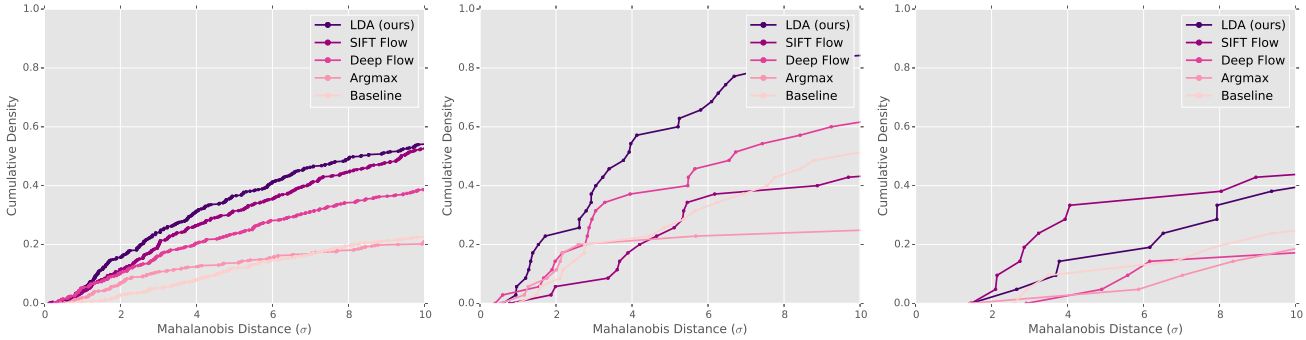

Figure 4. Comparison of sparse keypoint localization for our method, SIFT Flow [12] and Deep Flow [23]. The baseline measures the global alignment bias of the dataset (how well one would perform by simply assuming no flow). The argmax considers taking the single best match without regularization. The graphs measure the fraction of correspondences which fall within an increasing distance from groundtruth. 3 standard deviations is inperceptible from human annotator accuracy. From left to right: (a) aggregate results across all images, (b) the truck pair which our method localizes well, and (c) the biking pair for which our method fails to produce any meaningful correspondences.

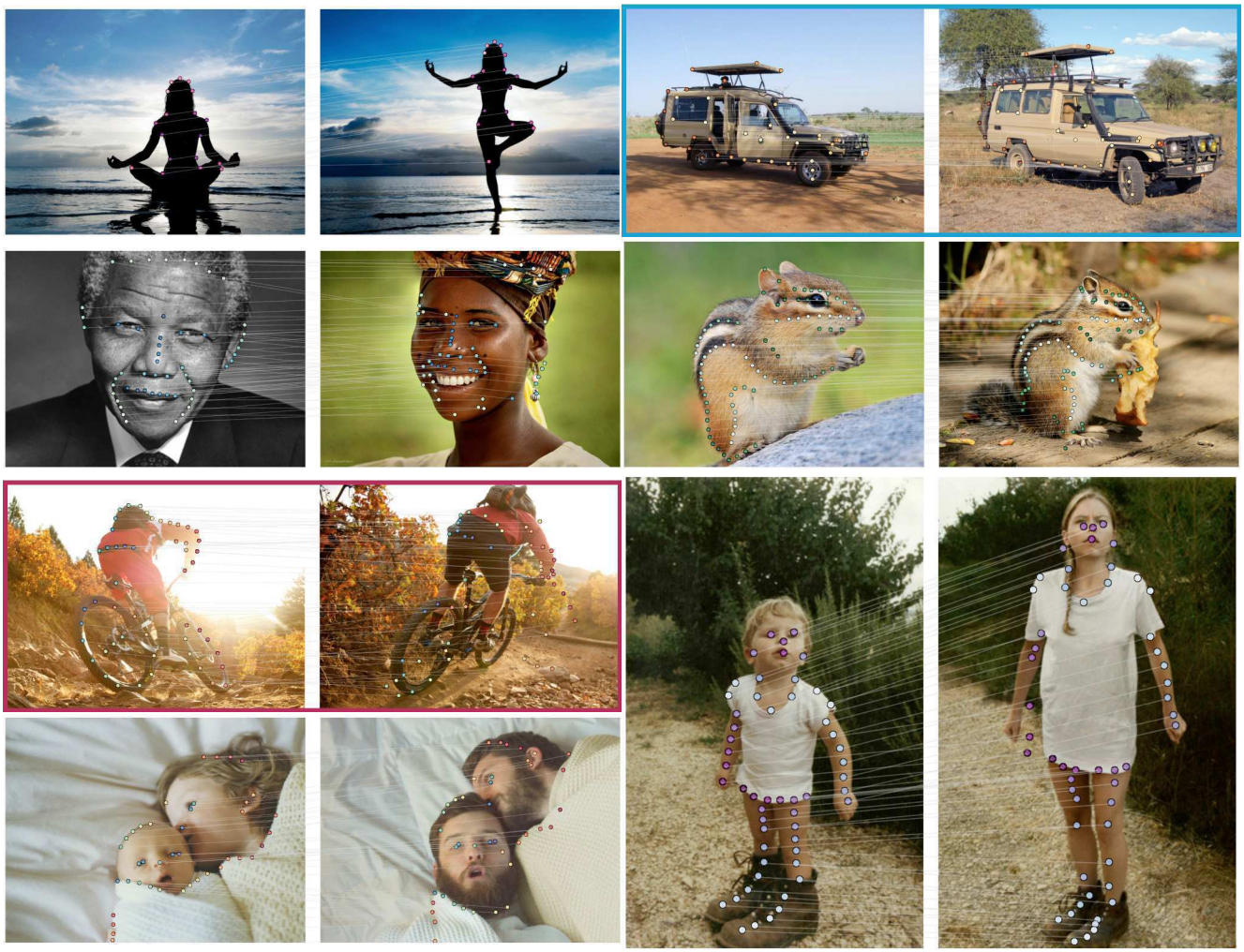

Figure 5. Example correspondences discovered by our method, across a broad range of image pairs from our dataset. The truck pair produces good localization of points (see Figure 4b), whilst the biking pair shows a failure to produce anything meaningful (see Figure 4c).