CAREER:

Beyond Edges: Visual Recognition with Adaptive Physically-based Representations

The last ten years have seen remarkable progress in machine vision with broad social and technical impact: the advent of the first real-time face detection systems, Kinect devices that precisely track limb positions, commercial product identification using Google Goggles. Yet these successes remain limited to narrow tasks; general scene understanding succeeds on fixed benchmarks but often fails in novel situated environments where users would like to deploy visual perception. Two key problems are an over-reliance on fixed, biased training data and simplified image representations that lack insight into the process of image formation and structure of the world.

Intellectual Merit: This project will address these two major underlying causes of the problem: data bias and weak image representations. The problem of **data bias** lies at the heart of the data-driven approach: the assumption that the distribution of features during deployment will remain unchanged from the training distribution.The PI of this proposal has been pioneering the development of adaptive methods for visual object recognition [63]. Her prior work involved learning transformations that adapt image representations to novel domains, based on a small amount of labeled data from the target domain. This work will form the foundation of the first major goal of the proposal: to investigate advanced adaptive image representations which generalize transform-based methods to a wider range of settings.

Modern recognition techniques have relied on image descriptors based on binned edge counts (SIFT, HOG, GIST, etc.) which treat all forms of contrast as equal—an edge is an edge is an edge—and almost completely ignore the process of image formation. These **weak representations** fail to capture the physics of light, shape and viewpoint. Recently, we advocated a new paradigm for probabistic physics-based vision and introduced a method which can infer physically accurate image content given only JPG values, significantly improving performance on certain tasks as well as a scheme for infering the true scene color given multiple color correspondences. The second major goal of this project is to infer and adapt calibrated descriptors from uncalibrated online imagery, leading ultimately to adaptive representations with a layered structure that factors the physical causes of the contrast.

By the end of our proposed effort we anticipate a unified model for adaptative perceptually-rich representations, such that discriminative (photometrically calibrated) local descriptors, and more generally layered representations, observed in one environment (e.g., a home) can be properly related to those collected in other environments (e.g., the web).

Broader Impacts: A benchmark database for adaptive perceptual representations will be created and disseminated, which aims to provide researchers with a common platform upon which to test novel photometrically rich adaptation techniques. Two new courses at UMass Lowell on computer vision and machine learning will be codesigned with perceptually-driven adaptive learning for real-world robotics as a central motivating scenario. Visual domain coursework modules will be created and made available to the community, and also will support UML's NSF-supported STREAM and Artbotics programs. The PI is committed to encouraging the participation of women in STEM education through direct and indirect means.

Keywords: computer vision; machine learning; domain adaptation.

CAREER:

Beyond Edges: Visual Recognition with Adaptive Physically-based Representations

1 Introduction

Imagine a future where robots perform household tasks, assist the elderly, respond to disasters and quickly learn new manufacturing and service skills. This seemingly futuristic scenario is set to unfold within our lifetimes! Recent advances in robotics have been stunning, with platforms such as robotic vacuums and quadcopters rapidly becoming ubiquitous and cost-effective. But to be truly useful, these machines will need a human-like ability to perceive and recognize objects in their environment.

Recognition of objects by machines has seen remarkable progress in the last decade, fueled by the ready availability of large image collections, the rapid growth of computational power, and advances in representations and algorithms. Commercial applications range from real-time face detection systems, now used in almost every consumer camera, to logo- and text-based product identification, including for example Google Goggles. Classification of images into a variety of complex categories has advanced as well: on the popular Caltech dataset of 101 categories (such as piano, panda, etc.,) classification accuracy improved four-fold, from 16% to nearly 65%, between 2004 and 2007. This inspired the creation of the much more difficult PASCAL VOC detection challenges during 2006-2012, with progress continuing to the point where average precision for many categories (dogs, cats, boats, etc.) in these challenging datasets is now better than 50%. The long-standing dream of machines that recognize objects is finally within reach!

To a large degree, the rapid gains have resulted from the adoption of data-centric approaches, and a slew of experiments have confirmed that learning from good and plentiful training data is quite powerful. But it is now becoming clear that we are reaching a plateau. For example, average precision for the important task of person detection in PASCAL VOC has increased from 42% to just 51.6% in the three years since 2008, in sharp contrast to the four-fold improvement on Caltech101 between 2004 and 2007. Poor generalization across datasets is also becoming evident: Torralba and Efros [67] recently reported that a person detector trained on the ImageNet database attains 59.6% precision on the test portion of that database, but loses one-third of that performance when tested on the LabelMe database, reaching only 39.0% precision.[1] These numbers point to a diminishing return on investment in the data-driven paradigm: training on ever increasing amounts of data is producing ever smaller gains in performance and generalization. Clearly, addressing this issue is key to making recognition accurate enough to be deployable on robotic platforms, such as those in homes, offices and hospitals.

The proposed research will address one of the major underlying causes of the bottleneck: the vicious cycle of weak image representations and data bias. The problem of **data bias** lies at the heart of the data-driven approach: the assumption that the distribution of features during deployment will remain unchanged from the training distribution. This certainly does not hold in the case of modern image representations, which do not model viewpoint, sensor noise, lighting, shadows and other

[1] Frustration with the current state of the art can by summarized by a quote from [67]: "(Recognition) will never work. We will just keep overfitting to the next data set."

source domain

target domain

train	test	SVM-bow	NBNN [7]
source	source	54 ± 2	61 ± 1
source	target	20 ± 1	19 ± 1

Figure 1: The data bias problem: Left: Edge-based representations of a "mug" classifer learned on a web image data set (the source domain) do not generalize well when applied to robot-collected images (target domain) where novel lighting and sensor conditions cause changes in edge distributions. Right: Performance of two object classification methods (an SVM over a bag-of-words representation (SVM-bow) and the Naive Bayes nearest neighbor (NBNN) classifier of [7]) degrades when trained on the web images and tested on the robotic image domain (bottom row).

scene variables, but rather learn 2D templates of edges (more on this later.) Any major change in these conditions with respect to the training conditions changes the distributions of these features and causes a drop in performance (see Figure 1.) Yet **most algorithms lack the ability to adapt to novel distributions or learn effectively from heterogeneous sources of data.** Recent advances in machine learning have begun to address this problem with domain adaptation techniques and the PI has been pioneering adaptive methods for visual object recognition [63]. Her prior work on adapting image representations to novel domains using transformations learned from a small amount of labeled data will form the foundation of **the first goal of the proposal: to investigate advanced adaptive image representations** which generalize transform-based methods to a wider range of settings.

While our first goal is to adapt representations to novel environments, regardless of the specific image features used, we must also re-examine these **weak representations**. Unfortunately, the potential of the data-driven approach is limited by the very image descriptors that have enabled its success so far. **Modern recognition techniques** have relied on image descriptors based on binned edge counts (SIFT, HOG, GIST, etc.; see Fig. 2, right) which **treat all forms of contrast as equal—an edge is an edge is an edge—and almost completely ignore the process of image formation.** These binned edge counts fail to capture the physics of light, shape and material properties, and fail to distinguish between edges due to e.g., cast shadows and those due to e.g., surface markings, the latter of which are much more likely to be important for identifying object categories. This forces recognition methods to require much more data to *learn* the variations caused by these irrelevant factors and, in turn, contributes to the problem of data bias.

Ultimately, our goal is to leverage descriptors which decompose local observations into perceptually meaningful components (see Figure 2, left), rather than treat appearance as a monolithic vector space to which machine learning is applied. The academic research community *has* pursued aspects of such representations, with principled, **physically-motivated methods** going back more than thirty years to Land [48], Horn [41], Barrow and Tennenbaum [2] and Marr [54], and other work referenced below, but these sorts of methods do not yet stand the test of real-world variety so they have not yet been useful in object recognition. A key problem is that these methods generally **require physically calibrated values as input**, such as the values linearly proportional to

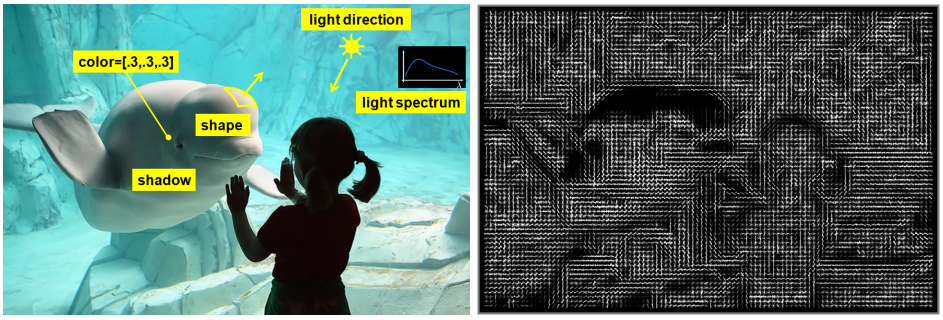


Figure 2: Recovering the interaction of light, surface color, shape and viewpoint will be essential for vision systems to succeed in the real-world, overcoming effects due to shadows, color cast, and complex illumination. To achieve this, they will need to move beyond current monochromatic gradient-based representations, such as the histograms of oriented gradients (HOG) shown on the right.

true scene radiance available in a camera's RAW file. Unfortunately, the data-driven approach has precluded the use of calibrated RAW data as most of the vast treasure trove of images stored online is in uncalibrated sRGB format. How can we relate the colors between the nonlinear sRGB web images and the linear sensor images shown in Figure 1 (source and target, respectively)? How can we compute physically-meaningful layers from the source images to train the next generation of descriptors?

Recently, we have advocated a new paradigm for probabilistic physics-based vision; we have introduced a method which can infer the radiometrically accurate RAW content given only sRGB values, significantly improving performance on certain tasks [73], as well as a scheme for inferring the true scene color given multiple color correspondences [61]. We propose to extend these methods such that they can be incorporated in contemporary recognition frameworks. **The second goal of this project is to infer and adapt calibrated descriptors from uncalibrated online imagery**, leading ultimately to adaptive representations with a layered structure that factors the physical causes of the contrast, as described in more detail below.

Our two goals are complementary: recovering scene parameters using photometric methods is a difficult, ill-posed problem and therefore we cannot rely on it entirely, while the data-driven approach of the first goal will allow us to adapt representations based on labeled data, and will also lend itself to adaptation between web images calibrated by our method and real-world images such as those obtained in specific environments (e.g, a household or office robot). By the end of our proposed effort we anticipate a unifed model for adaptative perceptually-rich representations, such that discriminative (photometrically calibrated) local descriptors, and more generally layered representations, observed in one environment can be properly related to those collected in other environments.

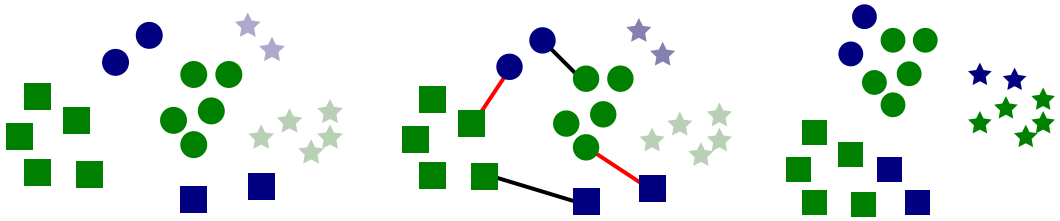
Both of the proposed goals will be pursued by an **integrated research and education program** at the University of Massachusetts, Lowell, where the PI will be appointed as Assistant Professor as of September 1, 2012. A **significant outreach initiative** towards underrepresented minorites will be a key focus of the proposed effort, including mentoring and engagement across the range of academic levels from high-school students to graduate researchers.

2 Background

Object recognition:. Early attempts at object recognition were focused on building 3D models and comparing them to 2D images by extracting geometric primitives (edges, lines, etc.) that are invariant to view and lighting changes [57, 56]. It became apparent, however, that such invariants cannot be extracted reliably in real-world images, and research has shifted toward learning-based approaches that use many training examples to *model* variation in viewpoint, illumination and category. Today, the prevalent approach involves extracting descriptor(s) from the image (or image patch, in the case of detection) and then applying a classifier based on k-nearest neighbors, neural networks with radial basis functions (RBFs), Fisher linear discriminants, support vector machines (SVMs), boosting algorithms, or something else. Significant recognition performance has been gained by further modeling changes due to pose and within-category variation using a constellation model or other parts-based approach, such as the recent LatentSVM [26] or poselets [8]; and precise pixel-level localizations are now often possible through simultaneous recognition and segmentation (e.g. [12]). In these approaches, most descriptors do not distinguish between underlying causes of image contrast. They treat all edges equally, and represent them with spatial histograms to gain insensitivity to changes in viewpoint and illumination. Figure 2 show an example of histogram of gradients (HOG) descriptors [19], and similar popular descriptors include SIFT [10, 50, 51], PCA-SIFT [44], gradient location-orientation histograms (GLOH) [55], spin images [49], shape context [3], locally binary patterns [3], steerable filters [32] and GIST [60].

Physical Layers: Instead of explicitly recovering scene variables, another approach is to use physics-based reasoning to decompose an image into separate "layers" that isolate particular physical effects. There are two flavors of these approaches: *photometric invariants* and *explicit decompositions*. Photometric invariants are deterministic transforms of pixel values—often simple ratios of intensities at different points and/or different color channels—that are relatively stable under certain changes in scene parameters, such as object shape and light direction [59, 34, 58], diffuse reflectance [72], cast shadows [29], or glossy reflectance [65, 76]. For some types of scenes, invariants can be a useful tool for discriminating between certain scene variables while eliminating sensitivity to other "distractors." Explicit decompositions, on the other hand, factor an image into its exact contributions from particular scene variables, and they usually require the solution of an ill-posed problem. Popular examples are decompositions according to albedo and shading, or "intrinsic images" [48, 66, 37], diffuse and specular reflectance [45, 64, 65, 52], scene colors and illuminant color (an aspect of color constancy) [11, 53, 31, 9, 30, 62, 69, 33, 13, 35], and cast shadows [28, 75, 38]. These operate by reasoning about each pixel's value in relation to its neighbors, often by analyzing small image patches. They continue to benefit from a transition to data-driven approaches, as exemplified by recent approaches to color constancy [33, 13, 35] monochromatic shadow detection [75], and reflectance/shading decomposition [66]. Most of the above methods require photometrically calibrated input images.

Adaptive Methods: Domain adaptation, or covariate shift, is a fundamental problem in machine learning, and has attracted a lot of attention in the machine learning and natural language community, e.g. [6, 20, 4, 43] (see [42] for a comprehensive overview.) Early visual adaptation methods were applied to domain shift in video, including work by Duan et al. [22], who proposed to adapt video concept classifiers (e.g. *person*, *vegetation*, *office*) between news videos collected from different news channels. To our knowledge, the first methods applied to visual category adaptation in still images are our earlier work [63, 47]. Other vision approaches for cross-domain



(a) Domain shift problem (b) Pairwise constraints (c) Transformed space

Figure 3: The key idea of our approach to domain adaptation is to learn a transformation that compensates for domain-induced changes. By leveraging (dis)similarity constraints (b) we aim to reunite samples from two different domains (blue and green) in the mapped space (c) in order to learn and classify new samples more effectively across domains. The transformation can also be applied to new categories (lightly-shaded stars). This figure is best viewed in color.

transfer include SVM-based methods: the method of [74] proposed an adaptive SVM, where the target classifier $f^T(\boldsymbol{x})$ is adapted from the existing, auxiliary classifier $f^A(\boldsymbol{x})$ via the equation $f^T(\boldsymbol{x}) = f^A(\boldsymbol{x}) + \delta f(\boldsymbol{x})$, where $\delta f(\boldsymbol{x})$ is the perturbation function. Domain transfer SVM [23] attempts to reduce the mismatch in the domain distributions, measured by the maximum mean discrepancy, while also learning a target decision function. A related method [21] utilizes adaptive multiple kernel learning to learn a kernel function based on multiple base kernels.

The disadvantage of [74, 23, 21] is the inability to transfer the adapted function to novel categories, which is limiting in object recognition scenarios, where the set of available category labels varies among data sets. In contrast, transform-based methods attempt to learn a perturbation over the feature space rather than a class-specific perturbation over the model parameters, typically in the form of a transformation matrix. Learning transformations has been an important problem in both the vision and machine learning communities (see [16, 39, 46, 15] for some additional examples). Other transform-based visual adaptation methods include [36, 24, 18].

3 GOAL I: Advanced Visual Domain Adaptation Methods

Suppose we have image or video data that we would like to run an off-the-shelf object detector on. Currently, no such off-the-shelf solution exists! With a few exceptions (e.g., face detection) most methods will not perform well unless we provide them with labeled examples of categories in our data. Rather than collecting a new data set each time, adaptive methods use existing labeled data sources to train a model, and then adapt it to the new data.

3.1 Goal I: Previous Results

The PI of this proposal has pioneered transformation-based domain adaptation methods for visual category learning. The key idea behind the work is to learn a transformation of the input feature space, such that target domain examples in the new representation are mapped "closer" (for some definition of closeness) to the source domain examples.

Adaptive Feature Transforms: Suppose that there are two domains $\mathcal{A}$ and $\mathcal{B}$ (e.g., source and target). Images from the source are embedded into a source vector space, and images in the target

into a target vector space. Given vectors $\boldsymbol{x} \in \mathcal{A}$ and $\boldsymbol{y} \in \mathcal{B}$, we learn a linear transformation W from $\mathcal{B}$ to $\mathcal{A}$ (or equivalently, a transformation W^T to transform from $\mathcal{A}$ to $\mathcal{B}$). Below, we will extend this model to learn non-linear transformations. If the dimensionality of the vectors $\boldsymbol{x} \in \mathcal{A}$ is d_A and the dimensionality of the vectors $\boldsymbol{y} \in \mathcal{B}$ is d_B, then the size of the matrix representing the transformation W is $d_A \times d_B$. We denote the resulting inner product similarity function between $\boldsymbol{x}$ and the transformed $\boldsymbol{y}$ as $\text{sim}_W(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T W \boldsymbol{y}$. We can think of this function as an inner product between $W\boldsymbol{y}$ (the feature vector $\boldsymbol{y}$ adapted by applying the mapping W) and $\boldsymbol{x}$. The goal is to learn the linear transformation given some form of supervision, and then to utilize the learned map in a classification or clustering algorithm. To avoid over-fitting, we choose a regularization function for W, which we will denote as $r(W)$ (choices of the regularizer are discussed below). Denote $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_A}]$ as the matrix of n_A training data points (of dimensionality d_A) from $\mathcal{A}$ and $Y = [\boldsymbol{y}_1, ..., \boldsymbol{y}_{n_B}]$ as the matrix of n_B training data points (of dimensionality d_B) from $\mathcal{B}$. Assume that the supervision is a function of the learned similarity values $\text{sim}_W(\boldsymbol{x}, \boldsymbol{y})$ (i.e., a function of the matrix $X^T W Y$), so a general optimization problem would seek to minimize the regularizer subject to supervision constraints given by functions c_i: $\min_W \; r(W)$, s.t. $c_i(X^T W Y) \geq 0, \;\; 1 \leq i \leq c$. Due to the potential of infeasibility, we can introduce slack variables into the above formulation, or write the problem as an unconstrained problem:

$$\min_W \; r(W) + \lambda \sum_i c_i(X^T W Y). \tag{1}$$

In both [63] and [47], we developed adaptation algorithms utilizing special cases of this model.

Nonlinear Case: In [63], we considered a special case of the transformation model where we assume that the dimensionality of the source is the same as that of the target (i.e., $d_A = d_B$). We focused on a particular regularizer and constraints that are a function of the learned Mahalanobis distances. Given two images $\boldsymbol{x}$ and $\boldsymbol{y}$ that are in the same class, or should be semantically similar, we constrain the distance $d_W(\boldsymbol{x}, \boldsymbol{y})$ to be small in (1) via the penalty $c_i(X^T W Y) = (\max(0, d_W(\boldsymbol{x}, \boldsymbol{y}) - u))^2$ for some upper-bound u; we define an analogous penalty for dissimilar images. The regularizer we employed is $r(W) = \text{tr}(W) - \log\det(W)$. Note that this regularizer can only be applied when the dimensionalities of the two domains are equal ($d_A = d_B$). This choice of regularizer and constraints has previously been studied as a Mahalanobis metric learning method, and is called *information-theoretic metric learning* (ITML) [46]. One of the key benefits of employing the LogDet regularizer $r(W) = \text{tr}(W) - \log\det(W)$ is that existing work has shown how to apply the resulting optimization algorithm for learning the transformation in kernel space. In other words, we can write the updates of the algorithm so that they only involve inner products between pairs of images. By replacing such inner products with arbitrary kernel functions, the resulting method learns non-linear transformations in the input feature space.

Asymmetric Transform: In [47] we extended the model of [63] to the more general case where the domains are not restricted to be of the same dimensionality, and where arbitrary asymmetric transformations could be learned. The fact that W is required to be symmetric and positive-definite may be overly restrictive for some applications. Positive definiteness implies that the learned transformation W can be factorized as $W = G^T G$, and thus the same transformation G is applied to both the source and the target. Further, the restriction that the dimensionalities of the source and target be the same is also potentially restrictive, for example when different features are utilized in the source and target. In order to avoid the restrictions of the ITML model for adaptation, we utilized an alternative regularizer that can generalize the model to use domains of differing dimensionalities but

still retains the benefits of kernelization. We focused on the particular regularizer $r(W) = \frac{1}{2}\|W\|_F^2$ and constraints that are a function of $\text{sim}_W(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T W \boldsymbol{y}$ for similar or dissimilar $\boldsymbol{x}, \boldsymbol{y}$ pairs. Our key technical contribution was to show that, for a large class of regularizers including the squared Frobenius norm above, the asymmetric transformation problem could be learned in kernel space, thus permitting non-linear transformations.

Our results demonstrated superior performance in challenging domain adaptation scenarios where a) test images come from a different sensor and lighting conditions than training images b) the categories utilized at test time are different from the ones at training time, and c) the image features utilized in the domains are of different dimensionality. Given this scenario, very few existing baselines can be applied—an SVM cannot be applied because all categories must be present at training time, and most other existing adaptation techniques assume that the source and target share the same underlying feature space. The results in [47] also demonstrated the benefits of learning a non-linear transformation: there is a significant improvement when using the kernelized version of our transformation model with a Gaussian RBF kernel.

3.2 Goal I: Proposed Research Effort

Despite the success of recent approaches to visual domain adaptation, it is clear that we are only beginning to understand the problem and offer effective solutions. One of the first limitations we would like to address is the scalability of transform-learning and discriminability of the resulting representation. Furthermore, we need to more fundamentally study the problem of separating data into domains in the case when the membership of the training data to source domains is unknown. We also propose to explore richer adaptation models that consider both global and *local* transformations, including category-specific transformations, and generalize well to novel categories.

Scalable Discriminative Transforms: While attractive, the approach in the previous section has two major flaws: First, unlike SVM parameter adaptation methods (see Section 2), transformation learning does not optimize the objective function of a strong, discriminative classifier directly; rather, it maximizes some notion of closeness between the transformed target points and points in the source. The second disadvantage is its increased computational complexity due to the high number of constraints, which is proportional to the product of the number of labeled data points in the source and target. This prevents the method from being applied to source domains with large numbers of points.

We plan to develop new algorithms which are able to discriminatively optimize transformations across multiple tasks, and generalize to new categories. We expect promising results from a method which jointly optimizes max-margin classifier parameters and replaces similarity constraints with more efficient hyperplane constraints to significantly reduce the training time of the algorithm, allowing it to be applied to a much larger number of source training points.

As shown in Figure 4, a jointly optimized method could handle the case of novel tasks by applying a task-independent transform W learned jointly on all tasks for which target labels are available, and obtain the adapted parameter for the new task, $W\theta^s$. This provides a way to adapt max-margin classifiers in a category-independent manner, by learning a shared transformation of the parameters. The main idea would be to learn the projection of source parameters into the target domain at the same time as learning the classifiers, using the same misclassification objective to optimize both the projection and the classifier parameters. Our proposed method would combine the strength of max-margin learning with the flexibility of the feature transform: because it operates

(a) Parameter adaptation of object with labels in target (*cup*)

(b) Adaptation cannot be applied to object without target labels (*phone*)

(c) Our method learns a task-independent transform W

Figure 4: State-of-the-art parameter-based adaptation techniques such as PMT-SVM [1] cannot generalize the learned domain shift to novel tasks/categories: (a) these methods learn a separate adapted parameter θ for each task by minimizing its distance from the source parameter θ^s learned on the source domain (images without borders) and minimizing misclassification of labeled images in the target domain (green border is positive, red is negative); (b) given a novel task with no labels in the target, these methods cannot predict the adapted parameter θ^s for that task; (c) in contrast, a jointly optimized transformation method learns a single parameter transform W for all tasks/categories and is thus able to predict $\theta = W\theta^s$ for the novel task.

over the input features, it can generalize the learned shift to novel categories in a way that parameter-based methods cannot.

As part of the proposed work, we will explore adaptive methods that jointly optimizes the SVM objective function while learning a transformation matrix from the target to the source. Let $x_1^s, \ldots, x_{n_A}^s$ denote the training points in the source domain, with labels $y_1^s, \ldots, y_{n_A}^s$. Let $x_1^t, \ldots, x_{n_B}^t$ denote the labeled points in the target domain, with labels $y_1^t, \ldots, y_{n_B}^t$. Our goal is to jointly learn affine hyperplanes that separate the classes in the source domain and a transformation from the points in the target domain into the source domain, such that the transformed target points lie on the correct side of the learned source hyperplanes. For a particular object category c we denote the normal to the affine hyperplane associated with that category as θ_c, and the offset of that hyperplane from the origin as b_c. As before, we will denote the transformation to be learned from target to source domains as W.

For simplicity of presentation, we show the optimization problem for single category classification with no slack variables. Our objective for the single category case is as follows:

$$\min_{W,\theta,b} \quad ||W||_F^2 + ||\theta||_2^2 \tag{2}$$

$$\text{subject to} \quad y_i^s \left(\begin{bmatrix} x_i^s \\ 1 \end{bmatrix}^T \begin{bmatrix} \theta \\ b \end{bmatrix} \right) \geq 1 \qquad \forall i \in \{1, \ldots, n_A\} \tag{3}$$

$$y_i^t \left(\begin{bmatrix} x_i^t \\ 1 \end{bmatrix}^T W^T \begin{bmatrix} \theta \\ b \end{bmatrix} \right) \geq 1 \quad \forall i \in \{1, \ldots, n_B\} \tag{4}$$

Note, this can easily be altered to add multiple classes by simply adding a sum over the regularizers on all θ_c parameters, and adding more constraints to represent those for each object class. Additionally, the formulation can be extended to use slack variables according to the standard soft-SVM technique. The objective function, written as in Equations (2)-(4), is not a convex problem

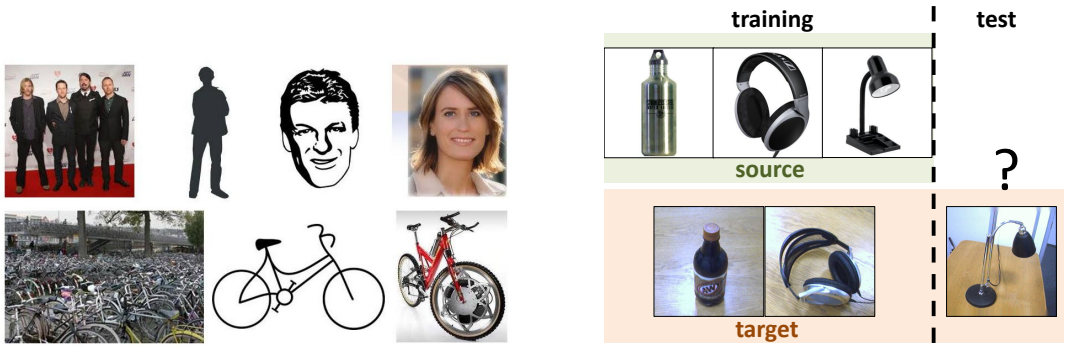


Figure 5: The need for richer adaptation models: (a) Training examples may contain several unknown source domains, such as these line drawings, close-up photos and far-away shots returned by web image search for *person* and *bicycle*. (b) A cross-domain transform learned using category-based constraints must generalize to categories not used to train it, such as the *lamp* category in the figure.

and so is both hard to optimize and does not have a global solution. Therefore, a standard way to solve this problem is to do alternating minimization on the parameters, W and (θ, b). We can effectively do this because when each parameter vector is fixed, the resulting optimization problem is convex.

Domain Discovery: While some first steps are being made towards addressing image data set bias, very little research has gone into what exactly constitutes a visual domain. Should every data set be treated as its own unique domain? Should similar data sets be grouped into a single domain? Could domain bias exist *within* a data set? We know domain bias exists, but we don't yet know how to detect it automatically.

In reality, one often has access to large amounts of labeled auxiliary data without the knowledge of the actual underlying domain of each example. For example, all images that can be retrieved through web search engines can either be thought of as one enormous "internet" domain, or as a collection of multiple domains. Fig. 5(a) shows that images returned by a search engine for the categories of "person" and "bicycle" can be separated into several *types*: close-up photos, photos taken from far away, shots with many people or many bicycles, line drawings, etc. Modeling such data as arising from multiple latent domains is an interesting problem that until now has been relatively unexplored.

In recent work to appear at ECCV12 we have demonstrated an approach that clusters the data into domains and shown results using a dataset of images returned by the Bing search engine ([5]). (Space limits preclude including additional discussion here, see [40] for details.) We propose to extend this line of work to address the key unresolved questions of domain discovery, including: What is a domain? Is the world a hierarchy of domains? Or is it a manifold of domains smoothly changing from one to another? Are supervised domains better than unsupervised domains? What is the optimal strategy for training on multiple data sets? How can unlabeled data be incorporated?

Transfer to Novel Categories: Most of the methods discussed so far adapt the representation based on training points available for a specific set of categories. In practice, the target domain is usually sparsely labeled (if at all), with only a small subset of categories available for training. However, we still want the learned representation to generalize to categories not used at training time, since we typically have training labels for them in the source domain. Figure 5(b) illustrates this problem. In

the current scheme, the domain transformation would be trained only on the categories with labels in the target, such as *bottle* and *headphones*, but would be applied to classify all categories, including *lamp*, which has no labels in the target. In previous results, transforms trained using purely category labels were less able to generalize to such novel categories. A possible reason is that the transforms did not capture the overall structure of the domain shift, but rather captured changes specific to the training categories.

Part of the research into richer models will include investigating regularizers that encourage transforms that capture both the category-specific and category-independent parts of the domain shift. Recall that the constraints used to optimize (1) encode *category*-level information, i.e., that the similarity $\text{sim}_W(\boldsymbol{x}, \boldsymbol{y})$ is high for $\boldsymbol{x}, \boldsymbol{y}$ pairs belonging to the same category, and low for all other pairs. Since the squared Frobenius norm regularizer does not impose any particular structure on W (such as symmetry for example) but simply encourages the values of W to be not too large, the learned transformation can be quite arbitrary and could over-fit to available constraints. In the extreme case, the learned similarity function could be one that outputs high values only if $\boldsymbol{x}, \boldsymbol{y}$ is nearly identical to one of the training constraint pairs.

We will investigate decomposing W into category-specific parts that are learned them jointly, regularized to be similar to each other. One way to proceed is by regularizing the rank of the a matrix W, constructed as follows: columns of W are vectorized versions of W_i. Since linear combinations of column vectors do not change rank, regularizing rank prefers to select W_i, such that most of them can be represented with a smaller set of bases (and each basis could be a column vector). In other words, we hypothesize that all transforms (category-specific ones) can be represented by a sparse set of basis transforms. Since rank of W is non-convex, we can use the nuclear-norm (trace-norm) of W as the proxy to the rank and use that as regularizer, which induces a convex optimization problem. It can be shown that we can still kernelize this problem efficiently, despite the issues that come up when introducing the vectorizing operation. We plan to test these ideas on existing benchmarks and novel multi-category data sets. More generally, we plan to investigate families of regularizers that satisfy the constraints of our transform-learning framework.

Timeline: In the first two years of our proposed effort, we will investigate the above extensions independently. In the third year, we will investigate how the models work with the representations developed in Goal II, as described in more detail below. In the fourth and fifth year, we will develop a unified model which encompasses results from both goals.

4 GOAL II: Adaptive Probabilistic Photometric Descriptors

We wish to develop image representations that go beyond traditional quantized edge histogram counts, and learn adaptive models that are sensitive to the photometric (and ultimately intrinsic) characteristics of surfaces and scenes. To do so requires first inferring physical scene properties, and then forming features for recognition. We first review our previous work addressing the former point, and then outline our proposed effort focusing on the latter.

4.1 Goal II: Previous Results

The PI of this proposal has pioneered the use of machine learning methods for learning richer image representations, recently developing new methods for multi-view color correspondence to learn the true color of a patch on an object [61] and for probabilistic color de-rendering to infer the RAW

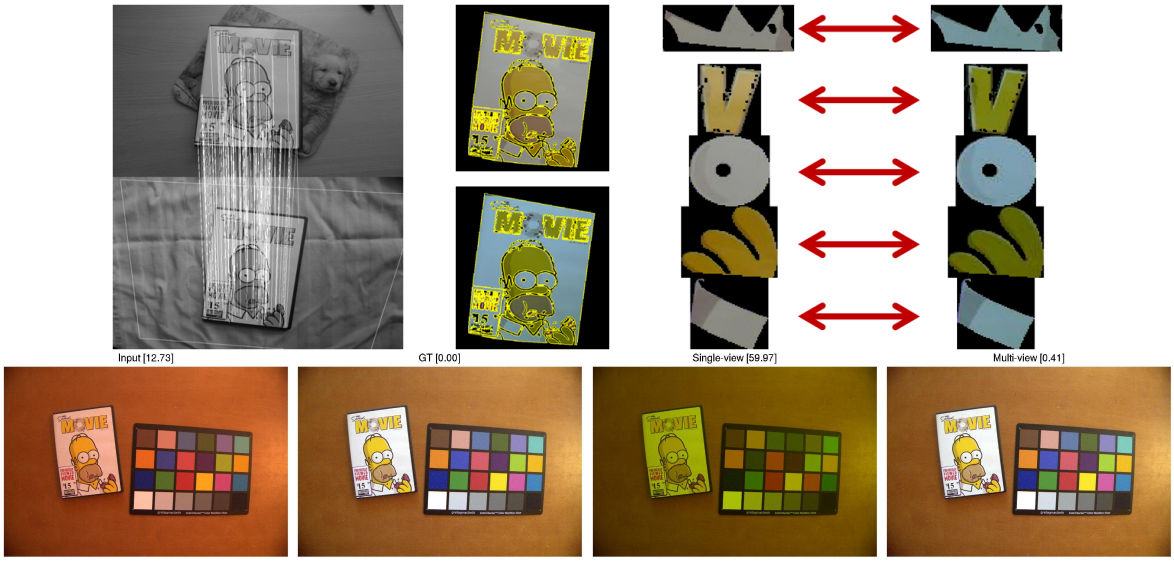


Figure 6: Color constancy with multi-view color correspondences [61]. (Top) corresponding color regions are extracted as follows: images are aligned and segmented using grayscale SIFT features, MSER are detected, the dominant color in each region is used to produce the final observed color correspondences. (Bottom) this row shows the original input image, and the illuminant-corrected images using the ground truth, the single-view spatial correlations method and the multiview spatial correlations method. The number in brackets is the angular error of the illuminant estimate w.r.t. the ground truth.

values from an uncalibrated sRGB (JPEG) image [73].

Color is known to be highly discriminative for many object recognition tasks, but is difficult to infer from uncontrolled images in which the illuminant is not known. Traditional methods for color constancy can improve surface reflectance estimates from a single images, but their output depends significantly on the background scene. In many recognition applications, such as on robotic platforms, we have access to image sets that contain multiple views of the same object in different environments; in this case correspondences between these images provide important constraints that can improve color constancy.

In [61] we introduced the multi-view color constancy problem, and presented a method to recover estimates of underlying surface reflectance based on joint estimation of these surface properties and the illuminants present in multiple images. The method can exploit image correspondences obtained by various alignment techniques including matching local region features. The key idea in our approach is to exploit correspondence constraints between multiple views when attempting color constancy. objects share the same underlying reflectance provides us with additional information about the illuminant in each scene. We combine the constraints implied by these shared reflectances with an existing single-view method for color constancy ([14]) to achieve our end goal.

Our results demonstrated that multi-view constraints can significantly improve estimates of both scene illuminants and object color (surface reflectance) when compared to a baseline single-view method [61]. Our method relies on having some number of regions matched across several views of the same object, such that each set of image regions corresponds to the same physical surface patch on an object. Such regions can be found using any number of multi-view techniques; one possibility is illustrated in Figure 6.

In [73], we introduced a new method for inferring calibrated RAW pixel measurements from uncalibrated JPEG imagery. Most digital images produced by consumer cameras and shared online exist in narrow-gamut, low-dynamic range formats. This is efficient for storage, transmission, and display, but it is unfortunate for computer vision systems that seek to interpret this data radiometrically when learning object appearance models for recognition. Indeed, most computer vision algorithms are based, either implicitly or explicitly, on the assumption that image measurements are proportional to the spectral radiance of the scene (called *scene color* hereafter), and when a consumer camera renders its digital linear color measurements to a narrow-gamut output color space (called *rendered color* hereafter), this proportionality is almost always destroyed. Fig. 7 shows an example. Existing methods attempt to undo these effects through deterministic maps to *de-render* the reported narrow-gamut colors back to their original wide-gamut sensor measurements. Unfortunately, deterministic approaches are unreliable, as this reverse narrow-to-wide mapping is one-to-many and has inherent uncertainty.

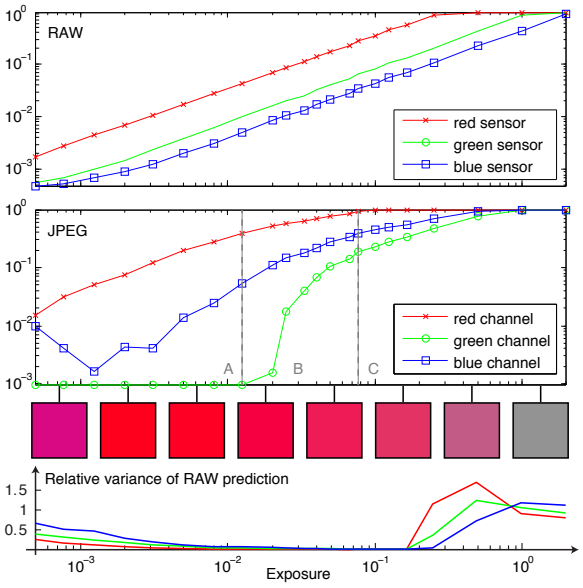


Figure 7: Probabilistic de-rendering [73]: RAW and JPEG values for different exposures of the same spectral scene radiance collected by a consumer digital camera (DMC-LX3, Panasonic Inc.), along with normalized-RGB visualizations of the reported JPEG colors at a subset of exposures. Apart from saturation, RAW values are linear in exposure and proportional to spectral irradiance; but narrow-gamut JPEG values are severely distorted by tone-mapping. Given only JPEG values, what can we say about the unknown RAW values—and thus the scene color—that induced it? How can we use all of the JPEG color information, including when some JPEG channels are saturated (regions A and C)? The proposed method answers these questions by providing a confidence level (bottom plot) of the estimate, which can be incorporated into radiometry-based computer vision systems.

Our GP-based approach overcomes these limitations, producing from each rendered (JPEG) color a probability distribution over the (wide gamut, high dynamic range) scene colors that could have induced it. The method relies on an offline calibration procedure involving registered RAW and JPEG image pairs, and from these it infers a statistical relationship between rendered colors and scene colors using local Gaussian process regression. This probabilistic approach provides a measure of confidence (e.g. the variance of the output distribution) for every de-rendered scene color, thereby eliminating the need for heuristic thresholds and making better use of the scene radiance information that is embedded in an Internet image.

Using a variety of consumer cameras we showed that these distributions, once learned from training data, are effective in simple probabilistic adaptations of two popular applications: multi-exposure imaging and photometric stereo. Our results on these applications were qualitatively better than using the corresponding deterministic approaches, especially in saturated and out-of-gamut regions. As supported by Fig. 7, one expects the one-to-many effect to be greatest near the edges of the output gamut (i.e., near zero or 255 in an 8-bit JPEG file), and practitioners try to

mitigate it using heuristics such as ignoring all JPEG pixels having values above or below certain thresholds in one or more of their channels. This trick improves the reliability of deterministic radiometric calibration, but it raises the question of how to choose thresholds for a given camera. ("Should I only discard pixels with values 0 or 255, or should I be more conservative?") A more fundamental concern is that this heuristic works by discarding information that would otherwise be useful. Referring to Fig. 7, such a heuristic would ignore all JPEG measurements in regions A and C, even though these clearly tell us *something* about the latent scene color.

Our GP-based approach overcomes these limitations, producing from each rendered (JPEG) color a probability distribution over the (wide gamut, high dynamic range) scene colors that could have induced it. The method relies on an offline calibration procedure involving registered RAW and JPEG image pairs, and from these it infers a statistical relationship between rendered colors and scene colors using local Gaussian process regression. This probabilistic approach provides a measure of confidence (e.g. the variance of the output distribution) for every de-rendered scene color, thereby eliminating the need for heuristic thresholds and making better use of the scene radiance information that is embedded in an Internet image.

4.2 Goal II: Proposed Research

We propose to extend the above methods to make them applicable to a wider range of conditions, and to incorporate them explicitly into representations for visual recognition. In the near term, we will generalize [73] to include a latent model of camera type, and to exploit a larger dataset of reference cameras, so that probabilistic de-rendering can proceed for images from unknown cameras (with the camera model being inferred from available image and/or meta-data, as appropriate). We also plan to investigate direct applications of [61], to other descriptors and to experimentally validate the utility of calibrated descriptors in practical recognition settings.

More generally, we plan to investigate how photometrically and chromatically calibrated low-level observations, such as may be provided by the above methods and their extensions, can be used to obtain physically-based representations that are useful for various recognition problems. We will investigate how photometrically calibrated representations can be integrated into several state-of-the-art recognition techniques. Our fundamental approach is to identify the underlying causes of appearance observed in an image, and use them in a representation for recognition.

We plan to develop an augmented variant of HOG/SIFT/GIST-style representations, where orientation bins are additionally labeled with intrinsic source type, or equivalently multiple redundant layers are represented. As described in the related work section, a number of approaches to factor physically-derived or "intrinsic" representations have been proposed, and generally have been demonstrated in laboratory environments. Extensions of our work above suggest such models should now be applicable more broadly, and that models can be learned from web data and applied in situated domains (leveraging as well the results from the previous goal.) Initial investigations in our lab using the method of Finlayson et al. [27] have suggested that an "Intrinsic-HOG" method could usefully distinguish edge energy due to shadows from those due to surface contrast. We propose to continue this investigation, and marry it explicitly with the work in [73] in the near term.

Conceptually, one could think of this mechanism as akin to adding a "depth" or "color" channel to a HOG representation, but where a physically-based cue is used instead. In contrast to classic depth and color extensions to these common representations, representations based on physical scene properties can provide a representation that directly addresses the underlying photometric

measures/physical causes of the scene. In our approach the different physics-based layers used by augmented descriptors can be thought of as separate observation channels. Various schemes for incorporating additional channels into existing classifiers will be considered, including linear and non-linear multi-kernel learning schemes. We will also consider both representations that group channels at a single pixel and those that represent them as separate planes, etc. We will investigate this scheme using several target recognition architectures, including: bag-of-word models on SIFT style descriptors (with chi-squared kernels, pyramid match and spatial pyramid match kernels, sparse variants [70]), deformable parts models using LatentSVM learning techniques, and poselet-style models and their extensions [71, 25] which overcome some of the traditional limitations of part-based appearance models. For each architecture, we will design extensions to existing algorithms to accommodate physics-based layer models, and experiment with different specific representations and optimization methods to find the most successful methods.

Existing domain adaptation methods, including our own present methods, will likely miss many of the most perceptually significant effects in the classes of rich representations suggested above. With few exceptions, methods for visual domain adaptation treat the perceptual signal as a homogeneous vector space, and attempt to find overly general transformations. While this appeals at the limit when one has infinite amounts of experience to learn from, it is naive to apply such methods to domains with known, specific constraints such as the interaction of (estimates of) lighting, shape, and viewpoint. A key long-term goal of our effort will be the extension of transformation learning to accommodate structured models with several constrained variables, possibly adopting the types of representations that have successfully been applied for supervised learning of structured representations, e.g., [17, 68]. While there are a number of hurdles that will have be overcome to develop equivalent models for the perceptual spaces we are interested in, we believe this to be a fruitful path of investigation in the long term.

Timeline: In the first two years of our effort we will investigate the basic combination of probabilistic de-rendering, color-constancy, and layered recognition models; we will exploit the domain adaptation methods from the previous goal as a pre- or post- processing step. In the third through fifth years we will investigate structured adaptation that is tightly integrated with the rich descriptors described above. By the final year, we expect to develop a unified model using a novel structured adaptation approach which will integrate techniques developed across both of our proposed major goals.

5 Dataset Curation and Dissemination of Results

One key additional broader impact of our work is the development of a benchmark database for adaptive perceptual representations, which aims to provide researchers with a common platform upon which to test novel photometrically rich adaptation techniques, and will hopefully make it easier for researchers to make continued progress in the field and to objectively evaluate our future contributions. We will continue our seminal effort for data set collection and curation, which has already had significant impact in the field. The collection in [63], introduced less than two years ago, has already been used and cited in numerous papers (including several oral presentations at major vision conferences by the PI and other authors). In addition, we will disseminate the results of our research broadly, via: distributions of code available for open use to the broader research community, publications in refereed conferences and journals, project web sites, as well as visiting lectures at universities and workshops.

6 Integrated Learning and Vision Curricular Development

As an integrated part of the above research goals, during the first two years of the proposed effort two new courses at UMass Lowell will be developed, on computer vision and machine learning, each with variants both at the graduate and undergraduate level to be taught in alternate years. These courses will be codesigned, so that the material is expressed in compatible notation and frameworks, in contrast to the traditional courseware available in machine learning and vision. These courses will provide a valuable service at UML CS, a PhD granting computer science department which has heretofore had no course in computer vision nor a course in machine learning. The topic of this proposal provides a unique opportunity to codesign these coures around paradigms of adaptive perceptual representations and real world robotics, which will be a significant motivator for many undergraduate students.

The proposed courses will also be developed in synergy with the existant proram in robotics education at UML, which is well-connected to the local robotics industry in the Route 128 area. The ability to provide sound, top-tier graduate education in vision and learning will be a significant broader impact and value to the local industry technology base.

As part of a broader outreach effort, we will incorporate visual domain adaptation into course-work modules that other instructors can include as perceptual components in their machine learning course, or as machine learning components in their computer vision course. (Or simply as extra material for courses led by other instructors who share the belief that the ideal way to present this material is with an equal balance between perception and learning.) we will present this material at short coures organized at major vision and learning conferences (most recently the PI co-led a CVPR11 tutorial), and plan to provide problem sets and courseware to accompany the material.

Perceptually adaptive learning can make classroom more exciting, and we plan to develop courseware that incorporates real-world robots with perception systems that can recognize objects based on web-based training and which can interact with students using KINECT-style gestures. Simplified versions of this material will also be integrated into K-12 outreach targeted at STEM underrepresented groups.

The educational material described above will also be incorporated into and will further support UML's STREAM and Artbotics programs, which are NSF-supported existing outreach efforts.

STREAM is a workshop to help K-12 educators utilize robotics as they teach STEM (science, technology, engineering, and math) subjects. This workshop will provide educators the opportunity to explore how they might use robotics in their own STEM instruction through interactive sessions as well as through presentations by other educators currently using robotics as a way to teach STEM. Representatives of local technology companies will also describe potential careers for students interested in robotics and STEM disciplines.

Artbotics is focused on learning about computer science, robotics, and art by creating interactive, public exhibits, and is a collaboration between UMass Lowell's Computer Science and Art departments with The Revolving Museum (located in downtown Lowell, MA). The program exists in many forms: a four-credit undergraduate course open to anyone at UMass Lowell, an after-school workshop for high school students, summer workshops for educators, and week-long summer camps for middle school students.

Through these and other venues, the PI of this proposal is committed to encouraging the participation of women in STEM education, via both direct mentoring and recruiting and indirect means by serving as a role model for excellence while maintaining a work/family balance.

References

[1] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 8

[2] HG Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*. Academic Press, New York, 1978. 2

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002. 4

[4] S. Ben-david, J. Blitzer, K. Crammer, and O. Pereira. Analysis of representations for domain adaptation. In *In NIPS*. MIT Press, 2007. 4

[5] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Neural Information Processing Systems (NIPS)*, December 2010. 9

[6] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *ACL*, 2007. 4

[7] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008. 2

[8] Lubomir Bourdev and Jitendra Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, pages 2–9, 2009. 4

[9] D.H. Brainard and W.T. Freeman. Bayesian color constancy. *JOSA A*, 14(7):1393–1411, 1997. 4

[10] Matthew Brown and David Lowe. Invariant features from interest point groups. In *In Proc. British Machine Vision Conf.*, 2002. 4

[11] G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980. 4

[12] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *Proc. IEEE International Conference on Computer Vision*, 2011. 4

[13] A. Chakrabarti, K. Hirakawa, and T. Zickler. Color constancy beyond bags of pixels. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 4

[14] A Chakrabarti, K Hirakawa, and T Zickler. Color constancy beyond bags of pixels. In *Proc. CVPR*, 2008. 11

[15] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Pattern Recognition and Image Analysis*, 2009. 5

[16] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. CVPR*, 2005. 5

[17] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002. 14

[18] Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. Translated learning: Transfer learning across different feature spaces. In *Proc. NIPS*, 2008. 5

[19] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 4

[20] H. Daume III. Frustratingly easy domain adaptation. In *ACL*, 2007. 4

[21] L. Duan, D. Xu, I. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *Proc. CVPR*, 2010. 5

[22] Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009. 4

[23] Lixin Duan, Ivor W. Tsang, Dong Xu, and Stephen J. Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009. 5

[24] Ali Farhadi and Mostafa Kamali Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV*, 2008. 5

[25] Ryan Farrell, Om Oza, Ning Zhang, Vlad I. Morariu, Trevor Darrell, and Larry S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *iccv*, 2011. 14

[26] Pedro F. Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–45, September 2010. 4

[27] G. Finlayson, M. Drew, and C. Lu. Entropy minimization for shadow removal. In *International Journal of Computer Vision*, 2009. 13

[28] G. Finlayson, S. Hordley, and M. Drew. Removing shadows from images. *Proc. European Conference on Computer Vision*, 2006. 4

[29] G.D. Finlayson and S.D. Hordley. Color constancy at a pixel. *Journal of the Optical Society of America A*, 18(2):253–264, 2001. 4

[30] G.D. Finlayson, S.D. Hordley, and P.M. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1209–1221, 2001. 4

[31] D. Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1):5–36, 1990. 4

[32] William T. Freeman and Edward H. Adelson. The design and use of steerable filters. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1991. 4

[33] P.V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp. Bayesian color constancy revisited. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 2008. 4

[34] T. Gevers and A. Smeulders. Color based object recognition. In *Image Analysis and Processing*, pages 319–326. Springer, 1997. 4

[35] A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *Image Processing, IEEE Transactions on*, 20(99):1–1, 2011. 4

[36] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *13th International Conference on Computer Vision 2011*, November 2011. 5

[37] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2335–2342, 2009. 4

[38] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011. 4

[39] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *CVPR*, 2004. 5

[40] J. Hoffman, K. Saenko, B. Kulis, and T. Darrell. Discovering latent domains for multisource domain adaptation. In *ECCV (to appear) preprint available at* `http://www.icsi.berkeley.edu/~saenko/pubs.html`, 2012. 9

[41] B. K. P. Horn. Obtaining shape from shading information. In *The Psychology of Computer Vision*, pages 115–155. McGraw-Hill, 1975. 2

[42] J. Jiang. A literature survey on domain adaptation of statistical classifiers. `http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/`. 4

[43] J. Jiang and C. X. Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June 2007. Association for Computational Linguistics. 4

[44] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Technical Report IRP-TR-03-15, School of Computer Science, Carnegie Mellon University and IntelRe- search Pittsburgh*, 2003. 4

[45] G.J. Klinker, S.A. Shafer, and T. Kanade. The measurement of highlights in color images. *International Journal of Computer Vision*, 2(1):7–32, 1988. 4

[46] B. Kulis, P. Jain, and K. Grauman. Fast similarity search for learned metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(12):2143–2157, 2009. 5, 6

[47] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011. 4, 6, 7

[48] E.H. Land and J.J. McCann. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1):1–11, 1971. 2, 4

[49] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using affine-invariant regions. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003. 4

[50] David Lowe. Object recognition from local scale-invariant features. In *In Proc. IEEE Intern. Conf. on Computer Vision*, 1999. 4

[51] David Lowe. Distinctive image features from scale-invariant keypoints. In *Intern. Journal of Computer Vision*, 2004. 4

[52] S. P. Mallick, T. Zickler, D. J. Kriegman, and P. N. Belhumeur. Specularity removal in images and videos: A PDE approach. In *Proc. European Conf. Computer Vision*, volume 1, pages 550–563, 2006. 4

[53] L.T. Maloney and B.A. Wandell. Color constancy: a method for recovering surface spectral reflectance. *JOSA A*, 3(1):29–33, 1986. 4

[54] D Marr. Representing visual information. In Allen R. Hanson and Edward M. Riseman, editors, *Computer Vision Systems*. Academic Press, New York, 1978. 2

[55] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005. 4

[56] J. Mundy. Object recognition in the geometric era: a retrospective. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward category-level object recognition*. Springer-Verlag, 2006. 4

[57] Joseph L. Mundy and Andrew Zisserman. *Geometric invariance in computer vision*. MIT Press, 1992. 4

[58] S.G. Narasimhan, V. Ramesh, and S.K. Nayar. A class of photometric invariants: Separating material from shape and illumination. In *Proc. IEEE International Conference on Computer Vision*, 2003. 4

[59] S.K. Nayar and R.M. Bolle. Reflectance based object recognition. *International Journal of Computer Vision*, 17(3):219–240, 1996. 4

[60] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001. 4

[61] T. Owens, K. Saenko, A. Chakrabarti, Y. Xiong, T. Zickler, and T. Darrell. Learning object color models from multi-view constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3, 10, 11, 13

[62] C. Rosenberg, T. Minka, and A. Ladsariya. Bayesian color constancy with non-gaussian models. In *Neural Information Processing Systems*. Citeseer, 2003. 4

[63] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010. 1, 2, 4, 6, 14

[64] P. Tan, S. Lin, L. Quan, and H.Y. Shum. Highlight removal by illumination-constrained inpainting. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2003. 4

[65] R.T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 178–193, 2005. 4

[66] M.F. Tappen, W.T. Freeman, and E.H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1459–1472, 2005. 4

[67] A. Torralba and A. Efros. An unbiased look at dataset bias. In *Proc. CVPR*, 2011. 1

[68] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. In *ICML*, 2004. 14

[69] J. Van De Weijer, T. Gevers, and A. Gijsenij. Edge-based color constancy. *Image Processing, IEEE Transactions on*, 16(9):2207–2214, 2007. 4

[70] J Wang, J Yang, K Yu, F Lv, T Huang, and Y Gong. Locality-constrained linear coding for image classification. In *cvpr*, 2010. 14

[71] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011. 14

[72] L.B. Wolff and J. Fan. Segmentation of surface curvature with a photometric invariant. *Journal of the Optical Society of America A*, 11(11):3090–3100, 1994. 4

[73] Y. Xiong, Saenko K., T. Darrell, and T. Zickler. From pixels to physics: Probabilistic color derendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3, 11, 12, 13

[74] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. *ACM Multimedia*, 2007. 5

[75] Jiejie Zhu, Kegan G. G. Samuel, Syed Z. Masood, and Marshall F. Tappen. Learning to recognize shadows in monochromatic natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 223–230, 2010. 4

[76] T. Zickler, S. P. Mallick, D. J. Kriegman, and P. N. Belhumeur. Color subspaces as photometric invariants. *Int. Journal of Computer Vision*, 79(1):13–30, August 2008. 4