World Wide Web (2016) 19:199–215
DOI 10.1007/s11280-015-0343-3

Active domain adaptation with noisy labels for multimedia analysis

Gaowen Liu[1] · Yan Yan[1] · Ramanathan Subramanian[2] · Jingkuan Song[1] · Guoyu Lu[3] · Nicu Sebe[1]

Received: 9 December 2014 / Revised: 25 February 2015 /
Accepted: 24 March 2015 / Published online: 12 May 2015

Abstract Supervised learning methods require sufficient labeled examples to learn a good model for classification or regression. However, available labeled data are insufficient in many applications. Active learning (AL) and domain adaptation (DA) are two strategies to minimize the required amount of labeled data for model training. AL requires the domain expert to label a small number of highly informative examples to facilitate classification, while DA involves tuning the source domain knowledge for classification on the target domain. In this paper, we demonstrate how AL can efficiently minimize the required amount of labeled data for DA. Since the source and target domains usually have different distributions, it is possible that the domain expert may not have sufficient knowledge to answer

✉ Yan Yan
yan@disi.unitn.it

Gaowen Liu
gaowen.liu@unitn.it

Ramanathan Subramanian
Subramanian.R@adsc.com.sg

Jingkuan Song
jingkuan.song@unitn.it

Guoyu Lu
luguoyu@udel.edu

Nicu Sebe
sebe@disi.unitn.it

[1] Department of Computer Science and Information Engineering, University of Trento, Trento, Italy

[2] Advanced Digital Sciences Center, UIUC, Singapore, Singapore

[3] Department of Computer Science, University of Delaware, Newark, DE, USA

each query correctly. We exploit our active DA framework to handle incorrect labels provided by domain experts. Experiments with multimedia data demonstrate the efficiency of our proposed framework for active DA with noisy labels.

Keywords Active learning · Domain adaptation · Noisy labels · Multimedia analysis

1 Introduction

In machine learning, supervised methods require sufficient labeled examples in order to learn a good model. However, it is difficult to acquire sufficient labeled data in many real world applications. Moreover, labeling is an intensive task requiring extensive human labor. In order to tackle this problem, several approaches have been proposed. *Semi-supervised learning* aims to exploit the consistency between labeled and unlabeled data for classification. *Active learning* (AL) focuses on selecting a small set of essential examples for querying labels from domain experts. *Domain adaptation* (DA, also called *Transfer learning*) facilitates classification when the training (source) and test (target) data are from different domains. Domain adaptation uses the knowledge acquired from a large number of labeled source examples and a few labeled target examples for classification in the target domain.

DA algorithms (see Pan and Yang [29] for a survey) seek to combine limited target data with the source data in order to adapt to the target domain. However, they typically tend to choose target examples *randomly* without considering which samples are most informative for classification in the target domain. Therefore, one question that needs to be examined is whether and how we can efficiently label target data for DA? Considering that the goal of both domain adaptation and active learning is to minimize labor-intensive data labeling, it would be worthwhile to integrate DA and AL in a single framework.

To our knowledge, very few works studied how to minimize the amount of labeled target data, especially under noisy labeling. A theoretical study on the number of labeled examples required to learn all targets to achieve an arbitrarily specified accuracy is presented in Yang et al. [44]. Two active transfer learning algorithms that allow for changes in all marginal and conditional distributions with the additional assumption that these changes are smooth are proposed in Wang et al. [37]. However, they do not consider noisy labels which are likely to occur in active DA scenarios. Shi et al. [33] propose active transfer learning, but their approach is limited by the unlikely assumption that identical prediction labels are generated for a target example by the out-of-domain (source) and in-domain (target) classifiers. Additionally, the error rate of the transfer classifier is not bounded, and only binary classification is considered here. Extending active transfer learning to multi-class classification as in this work, the upper-bound error rate increases considerably and consequently, the domain-adaptive classifier cannot classify correctly anymore.

In this paper, we investigate an adaptive DA algorithm within an AL framework able to cope with label noise. We also extend the binary classification to a multi-class classification problem through error-correcting output coding. We investigate how AL helps to minimize the numbers of labeled data for DA even under noisy labeling. Experiments on real-world datasets for headpose estimation and image classification demonstrate the efficacy of our proposed framework. To sum up, this paper makes the following contributions:

– An active domain adaptation framework under noisy labeling is proposed, and is shown to be effective for multimedia analysis;

– We integrate active learning with domain adaptation for a multi-class setting through error correcting output coding;
– The proposed framework is general, and potentially applicable to many multimedia problems.

The paper is organized as follows. Section 2 reviews related work from the perspective of active learning, domain adaptation and learning with noisy labels. Section 3 details active domain adaptation with noisy labels. Section 4 presents experimental results on headpose estimation and image classification, while Section 5 concludes the paper.

2 Related work

In this section, we review related work in the areas of active learning, domain adaptation and learning with noisy labels.

2.1 Active learning

Active learning (AL) involves asking the domain expert to label a small number of most-informative examples to facilitate classification. Based on query scenarios, AL can be divided into three types of settings: (i) Membership query synthesis, (ii) stream-based selective sampling and (iii) pool-based sampling. The pool-based scenario has been studied for many real-world problems in machine learning and computer vision. Uncertainty sampling is a common approach in AL. Distance from hyperplane for margin-based classifiers has been used as a measure of uncertainty in previous works. Tong and Koller [36] provided a theoretical motivation for SVM-based AL using the notion of a version space. Yan et al. [38] proposed a unified multi-class AL approach through error-correcting output coding based on the 'best worst case', which approximates the expected loss function with the smallest loss function among all the possible labels.

Hoi et al. [20] extended the Fisher information framework to the batch-mode setting for binary logistic regression. Sheng et al. [32] studied the problem of using several heuristics that take into account estimates of both oracle and model-uncertainty, and showed that data can be improved by selective repeated labeling. However, their analysis assumed both were equally and consistently noisy and annotation was a noisy process over some underlying true label. Liang and Grauman [25] introduced a novel criterion that requested a partial ordering for a set of examples that minimized the total rank margin in attribute space, subject to a visual diversity constraint.

Existing AL strategies can have uneven performance, being efficient on some datasets but ineffective on others, or inconsistent just between runs on the same dataset. Aodha et al. [2] proposed perplexity-based graph construction and a new hierarchical sub-query evaluation algorithm to combat this variability and to use the potential of expected error reduction. Elhamifar et al. [12] developed an efficient active learning framework based on convex programming, which can select multiple samples at a time for annotation. Unlike the state-of-the-art, their algorithm can be used in conjunction with any classifier type, including sparsity-based classifiers (SRC). Hua et al. [21] presented a collaborative computational model for AL with multiple human oracles. This approach leads not only to an ensemble kernel machine robust to noisy labels, but also to a principled label-quality measure detecting irresponsible labelers online.

Li and Guo [24] presented a novel multi-level AL approach to reduce the human annotation effort for training robust scene classification models. Different from most existing AL methods that can only query labels for selected instances at the class level, their approach established a semantic framework that predicted scene labels based on a latent object-based image representation, and was capable of querying labels at two different levels– the scene-class level and the latent object-class level. Yang et al. [46] proposed a semi-supervised batch mode multi-class AL algorithm for visual concept recognition. Chang et al. [7] proposed a novel convex, semi-supervised multi-label feature selection algorithm applicable to large-scale datasets.

2.2 Domain adaptation

Traditional machine learning algorithms are based on the assumption that training and test data share the same distribution in feature space. When the training and test distributions are different, the classification accuracy drops significantly. In such cases, domain adaptation (DA) between the two domains is desirable. DA assumes that the training and testing data could be from different domains and distributions. It is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems efficiently. The target of DA is to find some common property which is shared between the training (or source) and test (or target) domains.

Pan and Yang [29] identified three main research issues in DA: (i) what to transfer, (ii) how to transfer, and (iii) when to transfer. 'What to transfer' examines which knowledge can be transferred across domains or tasks. After discovering which knowledge can be transferred, learning algorithms are developed to describe the process of 'how to transfer'. 'When to transfer' studies the situations where the knowledge could be transferred, in order to guard against negative knowledge transfer that could hurt classification performance on the target domain.

There are several DA approaches. *Instance-transfer* involves re-weighting some source data for use in the target domain under the assumption that source data can be reused in the target domain [8, 22, 47]. *Feature-representation-transfer* attempts to find a 'good' feature representation that reduces the difference between the source and target domains as well as the classification/regression error [3, 9]. *Parameter-transfer* involves discovery of shared parameters or priors between the source and target models which can benefit from transfer learning [5, 13, 30]. *Relational-knowledge-transfer* builds a mapping of relational knowledge between the source and target domains [27].

In essence, transfer learning adapts useful source information to efficiently classify in the target domain whose attributes vary with respect to the source. Daume [9] proposed a feature replication method to augment features for transfer learning. Saenko et al. [31] and Kulis et al. [23] proposed a method for domain adaption using metric learning by generating cross-domain constraints. Dai et al. [8] used a boosting framework [14] to re-weight the importance of source and target samples for DA. Yao and Dorretto [47] extended the transfer boosting framework to include information from multiple sources. Yang et al. [43] adapted DA by learning a delta function between the source and target domains based on SVMs. This method seeks the target decision boundary which is close to the source decision boundary. Duan et al. [11] extended this method via multiple kernel learning by learning kernels that minimize the mismatch between source and target domains. Han et al. [19] proposed a framework for image attribute adaptation. Zhang et al. [48] proposed a DA framework for still-to-motion Adaptation (SMA) for human action recognition. Han et al. [18] proposed finding a low-dimensional optimal consensus representation from multiple

heterogeneous features for multi-view transfer learning. Han et al. [17] proposed a sparse multi-label learning method to circumvent the visually polysemous barrier of multiple tags.

2.3 Learning with noisy labels

Nowadays, with the exponential growth of user-generated web images and videos, there has been an increasing interest in learning models that can handle noisy labels for supervised learning. This is a practical problem due to the uncontrolled environments in which humans label data. Given the importance of learning from noisy labels, a great deal of progress has been made in this regard. Natarajan et al. [28] addressed the problem of risk minimization in the presence of random noise, and obtained generalizable results using unbiased estimators and weighted loss functions. Efficient algorithms were proposed with both methods with provable guarantees for learning under label noise. Yang et al. [45] proposed a multimedia retrieval framework based on semi-supervised ranking and relevance feedback. Yan et al. [41] proposed event-oriented dictionary learning for multimedia event detection. Biggio et al. [4] investigated the robustness of SVMs under adversarial label noise and proposed an improved method based on kernel matrix correction. Yan et al. [42] proposed a multi-task LDA method for multi-view action recognition.

In active learning, it is highly probable that the expert may have no information concerning some queries and cannot provide accurate labels. Du and Ling [10] studied AL under noisy labeling with a human-like oracle by introducing non-uniformly distributed noise. They made a realistic assumption that the less confident the oracle is in labeling the example, the larger is the effect of the noise. Sogawa et al. [34] proposed a pool-based active learning framework through robust measures based on density power divergence. By minimizing β-divergence and γ-divergence, one can estimate the model accurately even with noisy labels. Golovin et al. [15] tackled the fundamental problem of Bayesian active learning with noise, where they needed to adaptively select from a number of expensive tests in order to identify an unknown hypothesis sampled from a known prior distribution. Learning with noisy labels is especially important in DA scenarios. To the best of our knowledge, there is no work focusing on active transfer learning with noisy labels.

3 Active domain adaptation with noisy labels

Domain adaptation uses a small number of labeled samples from the target domain. However, taking into account that not all samples from the target domain are equally informative, an efficient sample selection strategy is preferable. To minimize the amount of labeled data in the target domain, we attempt AL using different sample selection strategies.

3.1 SVM-based domain adaptation

Recently, several adaptation methods for the support vector machine classifier (SVM) were proposed for video retrieval in Duan et al. [11]. In order to make the SVM classifier adaptive to a new domain, the target decision function $f^T(x)$ is formulated as:

$$f^T(x) = f^S(x) + \Delta f(x) \tag{1}$$

where x is the specific feature vector and $f^S(x)$ is the source decision function. $\Delta f(x)$ is the function of the mismatch between source and target domains.

Duan et al. [11] extended this method via multiple kernel learning. In this case, the target decision function is formulated as:

$$f^T(x) = \sum_{p=1}^{P} \gamma_p f_p(x) + \sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b \tag{2}$$

where $f_p(x)$ is the p-th pre-learned classifier trained using labeled data from both domains. P is the number of pre-learned classifiers. γ_p are the coefficients of the p-th pre-learned classifier. A linear combination of multiple kernels $\sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b$ is used to model $\Delta f(x)$ in this setting with a bias term b. M is the number of kernels and d_m are the coefficients of the m-th kernel. w_m^T is the transpose of the weight vector w_m and $\phi_m(x)$ is the nonlinear feature mapping function where base kernels can be calculated as $k_m(x_i, x_j) = \phi_m^T(x_i)\phi_m(x_j)$.

There are two objectives to minimize. The first objective is to reduce the mismatch between the source and target domains. Gretton et al. [16] proposed a similarity measure for two different distributions. The mismatch is measured by Maximum Mean Discrepancy (MMD) as in Huang et al. [22] based on the distance between the sample means from the source and target domains in the Reproducing Kernel Hilbert Space (RKHS) namely:

$$DIST(D^S, D^T) = \Omega(d) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(x_i^S) - \frac{1}{n_T} \sum_{i=1}^{n_T} \phi(x_i^T) \right\|_H \tag{3}$$

where x_i^S and x_i^T are the samples from the source and target domains, respectively. n_S and n_T are the number of samples in the source and target domains.

The second objective is to minimize the structural risk functional $J(d)$ in the target domain. If we combine these two objectives, the optimization problem is given by

$$\min_d G(d) = \frac{1}{2}\Omega^2(d) + \theta J(d) \tag{4}$$

where d is coefficient vector for the multiple kernels. $\Omega^2(d)$ is the distance between the source and target distributions. By introducing Lagrangian multipliers α, the dual form of the optimization is:

$$J(d) = \max_\alpha \alpha^T - \frac{1}{2}(\alpha y)^T \left(\sum_{m=1}^{M} d_m \widetilde{K_m} \right) (\alpha y) \tag{5}$$

This is equivalent to the dual form of SVM with kernel matrix $\sum_{m=1}^{M} d_m \widetilde{K_m}$, where $\widetilde{K_m}$ are the domain-adaptive rectified kernels. The optimization problem can be solved by an existing SVM solver, such as LIBSVM [6].

3.2 Multiclass active learning

Margin-based learning algorithms minimize the loss function $L(\cdot)$ with respect to the margin.

$$\min \frac{1}{m} \sum_{i=1}^{m} L(y_i f(x_i)) \tag{6}$$

Allwein et al. [1] proposed a unifying framework for studying the solution of multi-class categorization problems by reducing them to multiple binary problems. Firstly, we define a *coding matrix* $M \in \{-1, 0, +1\}^{k \times l}$. k is the number of classes and l is the number of binary classification problems. Let $M(r)$ denote the row r of M and $f(x)$ be the vector of predictions on an instance x, $f(x) = (f_1(x), ..., f_l(x))$. The basic idea is to predict with the label r, which row in $M(r)$ is the closest to the prediction $f(x)$, *i.e.*, predict label r that minimizes the distance $d(M(r), f(x))$.

Taking advantage of the confidence of binary predictions, Allwein et al. [1] proposed a loss-based decoding scheme. The idea is to choose the label r that is the most consistent with the predictions $f_s(x)$ in the sense that, if the example x was labeled r, the total loss on example (x, r) would be minimized over choices of $r \in Y$. The distance measure is the total loss on a proposed example (x, r).

$$d_L(M(r), f(x)) = \sum_{s=1}^{l} L(M(r, s) f_s(x)) \tag{7}$$

The predicted label $\hat{y} \in \{1, ..., k\}$ is:

$$\hat{y} = \arg\min_r d_L(M(r), f(x)) \tag{8}$$

Yan et al. [38] proposed an approximated sample selection strategy which uses the *best worst case* model to approximate the expected loss function with the smallest loss function among all the possible labels.

$$\arg\max_x \min_{y \in Y} \sum_{s=1}^{l} L(M(y, s) f_s(x)) \tag{9}$$

If y_x is the predicted label for example x, Eq. (9) becomes:

$$\arg\max_x \sum_{s=1}^{l} L(M(y_x, s) f_s(x)) \tag{10}$$

Here, we choose the most ambigous examples with the maximum expected loss for the predicted label.

3.3 Modeling with noisy labels

Information-theoretic methods can be used to model expert labeling knowledge. In the traditional AL scenario, the expert is able to provide a label for each queried instance. Then, the objective of uncertainty sampling based AL is to query the instance with the highest entropy. We model the domain expert as either knowledgeable to label an instance or not knowledgeable. The Knowledge Base (N) is defined as the union of instances (N^+) which have been labeled by the domain expert, and those instances (N^-) which the domain expert is unable to label.

The expected entropy of an unlabeled instance x_i with respect to sets N^+ and N^- is given by:

$$E = P(x_i \in N^+)E(y_i|x_i \in N^+) + P(x_i \in N^-)E(y_i|x_i \in N^-)$$

where $E(\cdot)$ is the entropy of samples x_i with respect to the predicted classifier label. Moreover, in the above equation $E(y_i|x_i \in N^-) = 0$ due to the definition of conditional entropy. The diverse density concept proposed in Maron and Lozano-Perez [26] is adopted to estimate $P(x_i) \in N^+$.

3.4 Framework

Considering that the goal of both DA and AL is to minimize intensive data labeling, it is reasonable to investigate how combining them can further minimize data labeling on the target. We propose an active DA under noisy labeling framework as shown in Figure 1. We use labeled source, labeled and unlabeled target data to train the transfer classifier. Then, we use AL to select unlabeled target data to be labeled by the expert, and add the same to labeled target data to update the transfer classifiers.

Algorithm 1 presents the active DA under noisy labeling algorithm. We initially randomly select one sample per category. Steps (4-8) represent the DA procedure. We combine labeled target samples D_l^s with labeled source samples D_l^t to train an adaptive SVM classifier $f^{T^m}(x)$ on the target domain D^t. To this end, we employ alternative coordinate descent to optimize variables α and d in Eq. (5). η_t is the learning rate and g_t denotes the update direction. We iterate this procedure T_{max} times. Steps (9-12) represent the AL procedure. In step (9), we calculate loss values for all the unlabeled target samples. We choose those unlabeled target samples that produce the least loss to be labeled by experts, and then add these samples to the labeled target domain. Steps (13-19) represent the procedure adopted to deal with noisy labels. If the expert

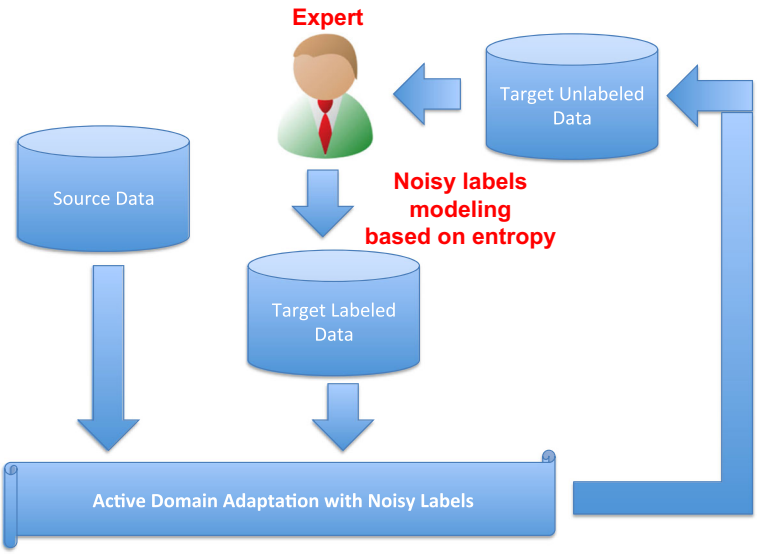

Figure 1 Proposed framework for active domain adaptation with noisy labels

does not know the label for x_i, the algorithm will include x_i in the negative knowledge base (N^-). Step (19) is to update the knowledge base N. We iterate this procedure K times.

Algorithm 1 Active Domain Adaptation under Noisy Labeling.

Input: Labeled *source* data D^s and unlabeled *target* data D^t. Let $D^t = D_l^t \cup D_u^t$. Randomly label one *target* sample per class and add them to D_l^t.
Output: Target sample label.

for $k = 1, ..., K$ do
  Perform domain adaptation on D^t using samples from $D_l^s \cup D_l^t$ to obtain $f^{T^m}(x)$.
  • **for** $t = 1, ..., T_{max}$ do
    • Solve dual SVM variable α_t using LIBSVM with kernel matrix $\sum_{m=1}^{M} d_m \widetilde{K_m}$.
    • Update the base kernel coefficients d_t by $d_{t+1} = d_t - \eta_t g_t$.
  • **end for**
  For all the samples $x_i \in D_u^t$, calculate loss function $\arg\max_x \sum_{s=1}^{l} L(M(y_x, s) f_s(x))$.
  For all the samples $x_i \in D_u^t$, estimate $P(x_i) \in N^+$, then calculate expected entropy of x_i.
  Select samples s^* according to the sum of least loss and expected entropy.
  Get label y_{s^*}.
  if the expert *knows* the label **then**
    Add sample $s^* = (x_{s^*}, y_{s^*})$ to D_l^t.
    $N^+ \leftarrow N^+ \cup x_i$.
  else
    $N^- \leftarrow N^- \cup x_i$.
  end if
  $N \leftarrow N^- \cup N^+$ and update knowledge N.
  Classification using $f^{T^m}(x)$ on target domain test data.
end for

4 Experiments

In this section, we test the proposed active DA method for cross-domain headpose estimation (proposed earlier in Yan et al. [39, 40]) and cross-domain web image classification (proposed in Saenko et al. [31]).

4.1 Cross-domain headpose dataset

In video surveillance, knowing *where a person is looking at* is important. However, headpose estimation or classification from surveillance videos can be very hard, due to the low resolution and noise characterizing the sensor data. We focus on headpose estimation from low-resolution images acquired using a multi-camera system.

The CLEAR 2007 dataset [35] illustrated in Figure 2a provides multi-view images, output from four cameras placed in the room's corners. This dataset includes 15 persons rotating in-place, and exhibiting all possible head orientations while wearing a magnetic motion sensor (flock-of-birds) to measure their head pose. The task is to estimate the person's 3D head orientation with respect to the room's coordinate system, and to obtain a robust, joint pose estimate from all four views instead of employing only a single camera view for analysis.

In order to evaluate cross-domain headpose classification, we used the DPOSE dataset (described in Rajagopal et al. [30]) shown in Figure 2b. DPOSE is recorded under the same settings as CLEAR, with both static and moving persons (only data corresponding to static persons are used in our experiments). As evident from Figure 2, the illumination and recording environments are very different in the CLEAR and DPOSE datasets.

We firstly localize the head in each of the four views using the procedure described in Rajagopal et al. [30]. The localized head regions are then resized to 20×20 resolution. We then concatenate the head crops from the four views on which visual features are extracted. Head pan is divided into eight classes, each denoting a 45° pan range, and for each head pan range, the tilt is quantized into three classes– namely *frontal* [−20°, 20°], *upward* (20°, 90°] and *downward* (−20°, −90°]. This leads to a total of 24 headpose classes (*e.g.* pan range (-22.5,22.5) with *frontal*, *upward* and *downward* tilts denote headpose classes 1–3). We divide the 4-view head image into 25 patches (every patch is 4×4). For the visual features computed over each view, we use HOG (81 dimensions) and skin pixel histograms (25 dimensions denoting the number of skin pixels in each patch). Then, we concatenate these features to derive a 106-dimensional vector per view, and a 424-dimension vector over the 4-view image.

We use several baseline methods to evaluate and compare our transfer learning results. $S_A T_B$ means we train on source domain A and test on target domain B. $S_B T_B$ means we train on target domain B and test on B. $S_{(A+B)} T_B$ means we train on both A and B and test on B. $TrAdaboost$ means we use the Adaboost algorithm [14] trained on labeled source and target data. AMKL_random means we use adaptive multiple kernel learning and randomly label target samples. AMKL_active (our method) means we use AMKL and actively label the target samples. For all the experiments, we report the mean accuracy on 5 randomly selected train/test sets. SVM parameter $C = 1$ in all the experiments. We use 100 images per class in the source domain and query 24 samples (one sample/class) to label every round. To begin with, there are 100 unlabeled images per class in the target domain.

Figure 3 compares classification accuracies achieved using various approaches over 30 rounds of active learning. Evidently, we can see that our active transfer learning algorithm outperforms all the considered baselines. Clearly, our method efficiently learns about the target domain upon incorporating knowledge from a few target examples. Also, employing information from all four camera views achieves superior performance as compared to monocular analysis. Comparing AMKL_active with AMKL_random, we see that in both the monocular and multi-view cases, our approach outperforms AMKL_random after 10 rounds

Figure 2 4-view exemplar from the (**a**) CLEAR and (**b**) DPOSE datasets. Automatically extracted face crops are shown on the bottom right inset

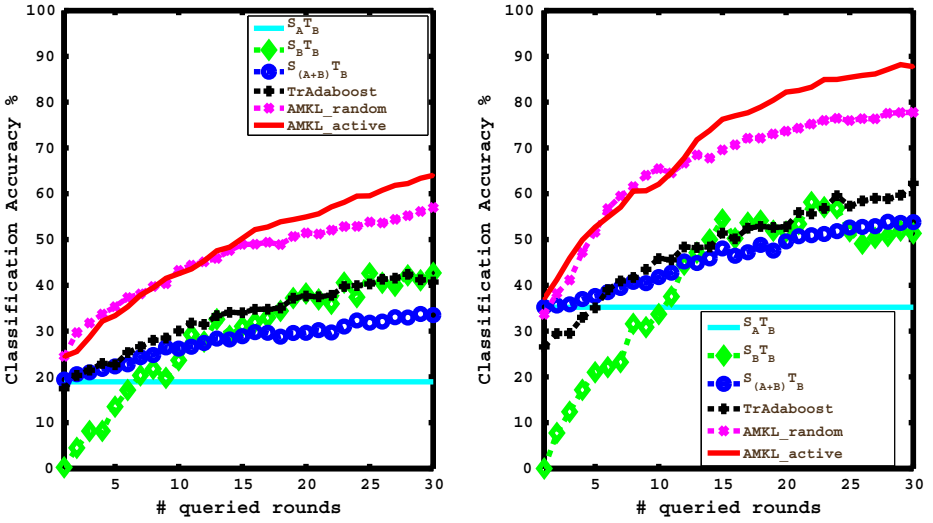

Figure 3 Classification accuracies with (*left*) single view and (*right*) 4 views

of AL, and the benefit of learning from the most informative samples is reflected in the fact that AMKL_active outperforms AMKL_random by more than 10 % after 30 rounds while classifying with 4-view information.

Figure 4 shows the confusion matrix over 24 headpose classes using active transfer learning after 30 rounds. We can conclude that most of the target samples are correctly classified. Moreover, most of the misclassified samples belong to nearby classes, which means the samples are only misclassified with respect to head tilt, while the head pan is classified correctly. This again demonstrates the robustness of our active DA framework for headpose estimation. We also evaluate the effects of using five different types of loss function in the AL module– (i) Logistic loss $1/(1 + e^{2yf(x)})$, (2) Exponential loss e^{-x}, (3) Hinge

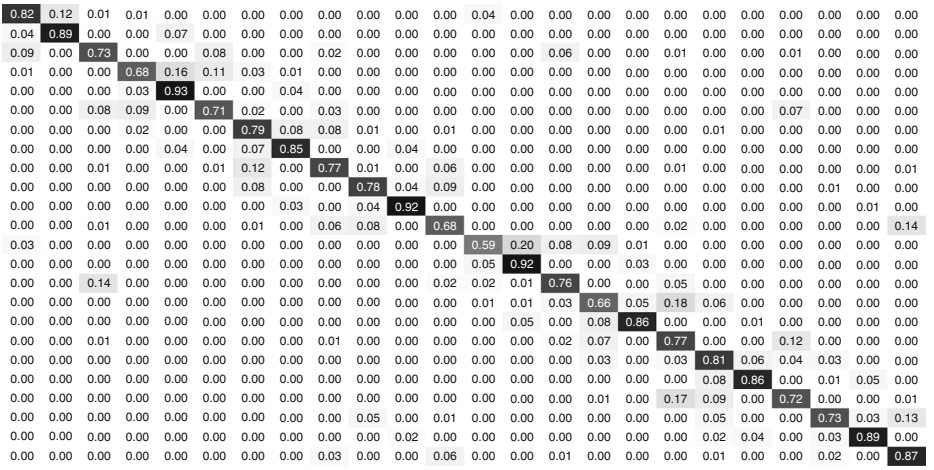

Figure 4 Confusion matrix over 24 classes for active DA after 30 rounds

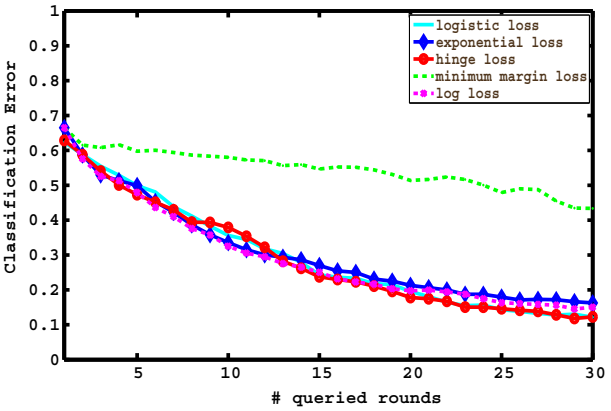

Figure 5 Evaluating active DA classification error with different loss functions

loss $(1 - y)_+$, (4) Minimum margin loss e^{-100x} and (5) log loss $log(1/(1 + x))$. Figure 5 presents the active transfer learning classification error obtained on these different loss function. We observe that hinge loss achieves the better performance among all loss functions, which implies that active transfer learning works optimally if identical loss functions are employed in the DA and AL modules.

Since querying sample labels for AL can also be done in a batch mode, we examine the extent of reduction in classification error for varying number of queried samples at every round. Figure 6 shows the progressive reduction in classification error with differing number of queried samples (4, 8 and 12 samples/class/round) for AL. From Figure 6, we can see that the classification accuracy is not influenced much by varying the number of queried samples per round. However, choosing a moderate number of queried samples per round appears to be optimal since the error is minimal when 8 samples per round are queried as compared to querying 4 or 12 samples per round. Finally, we evaluate the robustness of our active DA framework to noisy labels. Figure 7 compares classification accuracies achieved

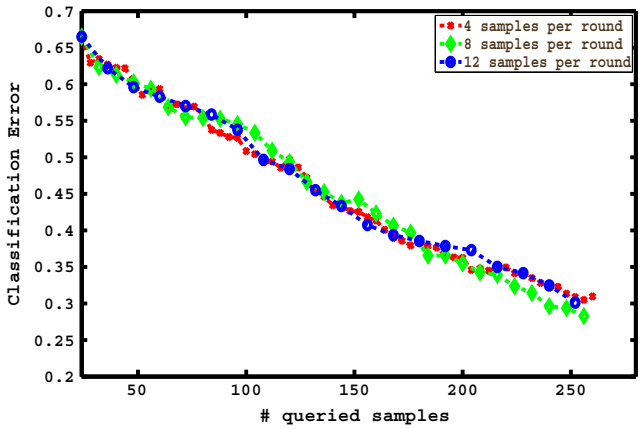

Figure 6 Evaluating our active DA framework with batch mode querying by varying number of queried samples/class/round

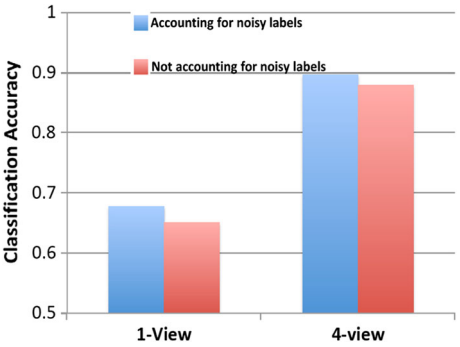

Figure 7 Evaluating active domain adaptation with noisy labels modeling strategy

with and without modeling for noisy labels in the AL module (steps 13–19 in Algorithm 1). Note that about 3 % higher accuracy is achieved by accounting for noisy labels when using both monocular and 4-view image features.

4.2 Cross-domain Berkeley web image dataset

The Berkeley image dataset consists of three types of images: web images (from amazon), images from a digital SLR camera (high resolution image), and low-resolution webcam images, as shown in Figure 8. Each domain has 31 categories of images. While the digital SLR camera and webcam images capture the same objects, the viewpoint and image resolutions are different.

Our objective on the Berkeley dataset is to perform object recognition across image domains. For all the experiments, we report mean accuracy obtained on 5 randomly selected train/test sets. SVM parameter $C = 1$ in all the experiments. For each object category,

Figure 8 Exemplars from the Berkeley web image dataset. (from *top to bottom*) Web (amazon), digital SLR camera (high resolution image) and webcam (low resolution image)

Table 1 Source domain - *webcam* images

	webcam→dslr	webcam→amazon
$S_A T_B$	0.19 ± 0.02	0.09 ± 0.01
$S_B T_B$	0.37 ± 0.01	0.18 ± 0.02
$S_{(A+B)} T_B$	0.28 ± 0.02	0.15 ± 0.01
Saenko et al. [31]	0.27 ± 0.02	0.19 ± 0.01
TrAdaboost (Dai et al. [8])	0.25 ± 0.02	0.17 ± 0.02
DA	0.35 ± 0.02	0.20 ± 0.01
ADA	0.61 ± 0.02	0.23 ± 0.01
ADAN	$\mathbf{0.65 \pm 0.02}$	$\mathbf{0.27 \pm 0.02}$

The bold means the largest values in the columns of tables

there are a small number of labeled samples in the target domain (3 in our experiment). For the source domain, we use 8 labels per category for *webcam/dslr* and 20 for *amazon*. As low-level visual descriptors, we use the pre-compute SURF features. A codebook of size 800 is constructed by k-means clustering. We firstly normalize the feature vector and then repeat the experiment as in Saenko et al. [31]. Descriptions of the several baseline methods compared are as follows:

- $S_A T_B$ - We train on source domain A and test on target domain B.
- $S_B T_B$ - We train on target domain B and test on B.
- $S_{(A+B)} T_B$ - We train on both A and B, and test on B.
- Saenko et al. [31] - A metric learning-based DA approach.
- $TrAdaboost$ [8] - DA based on the *Adaboost* algorithm.
- DA - DA with adaptive multiple kernel learning (AMKL) and randomly label target samples.
- ADA - DA and actively label target samples.
- ADAN - Proposed DA method accounting for noisy labels.

Tables 1, 2 and 3 compare classification accuracies achieved with the different approaches when trained on images from the webcam, dslr and amazon domains respectively. We make the following observations from these tables: (i) Superior performance

Table 2 Source domain - *dslr* images

	dslr→webcam	dslr→amazon
$S_A T_B$	0.15 ± 0.01	0.04 ± 0.01
$S_B T_B$	0.40 ± 0.03	0.18 ± 0.02
$S_{(A+B)} T_B$	0.20 ± 0.02	0.08 ± 0.01
Saenko et al. [31]	0.31 ± 0.03	0.15 ± 0.02
TrAdaboost (Dai et al. [8])	0.44 ± 0.03	0.10 ± 0.02
DA	0.49 ± 0.02	0.15 ± 0.02
ADA	0.59 ± 0.02	0.22 ± 0.02
ADAN	$\mathbf{0.63 \pm 0.02}$	$\mathbf{0.31 \pm 0.02}$

The bold means the largest values in the columns of tables

Table 3 Source domain - *amazon* images

	amazon→dslr	amazon→webcam
$S_A T_B$	0.04 ± 0.02	0.08 ± 0.01
$S_B T_B$	0.36 ± 0.03	0.38 ± 0.02
$S_{(A+B)} T_B$	0.10 ± 0.03	0.14 ± 0.02
Saenko et al. [31]	0.32 ± 0.02	0.48 ± 0.03
TrAdaboost (Dai et al. [8])	0.22 ± 0.03	0.38 ± 0.01
DA	0.28 ± 0.01	0.39 ± 0.02
ADA	0.36 ± 0.03	0.45 ± 0.01
ADAN	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.49 \pm 0.03}$

The bold means the largest values in the columns of tables

is always achieved using S_B as compared to S_A, which proves the need for DA for object recognition on the Berkeley dataset. (ii) While the inductive TrAdaboost and metric learning-based DA approaches perform favorably with respect to $S_{(A+B)} T_B$, they are generally outperformed by the AMKL-based DA approaches studied in this work. (iii) ADA outperforms DA considerably, implying that AL greatly benefits DA for object recognition. (iv) ADAN outperforms ADA by up to 5 % on an average, implying that our approach which explicitly accounts for label noise greatly benefits AL. (iv) ADAN consistently produces the best recognition performance demonstrating the efficiency of the proposed active DA framework.

Commenting on the computational time required for our proposed algorithm, model training for cross-domain multi-view headpose estimation and object recognition required 20 minutes with cross-validation on a workstation with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz $\times$ 17 processors implying that our algorithm can be applied on large-scale datasets.

5 Conclusion

We propose an active transfer learning framework which explicitly accounts for ambiguous labels provided by the domain expert. We also extend traditional active learning for binary classification to a multi-class setting through error-correcting output coding. Extensive experiments on cross-domain multi-view headpose estimation and object recognition demonstrate the effectiveness of our proposed method. In particular, the ability to select the most informative samples for active learning and handle label noise improves classification performance with respect to random sample selection. Developing DA approaches that (i) incorporate useful information from unlabeled target samples and (ii) learn from multiple sources will be the focus of future work.

Acknowledgments This work was partially supported by the MIUR Cluster project Active Ageing at Home, the EC project xLiMe and A*STAR Singapore under the Human-Centered Cyber-physical Systems (HCCS) grant.

References

1. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. JMLR **1**(1), 113–141 (2000)
2. Aodha, O.M., Campbell, N.D., Kautz, J., Brostow, G.J.: Hierarchical subquery evaluation for active learning on a graph. In: CVPR (2014)
3. Argyrious, A., Evegenious, T.: Multi-task feature learning. In: NIPS (2007)
4. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. J. Mach. Learn. Res. **20**, 97–112 (2011)
5. Bonilla, E., Chai, K., Williams, C.: Multi-task gaussian process prediction. In: NIPS (2008)
6. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001)
7. Chang, X., Nie, F., Yang, Y., Huang, H.: A convex formulation for semi-supervised multi-label feature selection. In: AAAI (2014)
8. Dai, W., Yang, Q., Yu, Y.: Boosting for transfer learning. In: ICML (2007)
9. Daume, H.: Frustratingly easy domain adaptation. In: ACL (2007)
10. Du, J., Ling, C.X.: Active learning with human-like noisy oracle. In: ICDM (2010)
11. Duan, L., Xu, D., Tsang, I.W.: Visual event recognition in videos by learning from web data. In: CVPR (2010)
12. Elhamifar, E., Sapiro, G., Sastry, S.: A convex optimization framework for active learning. In: ICCV (2013)
13. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: SIGKDD (2004)
14. Freund, Y., Schapire, R.: A short introduction to boosting. J. Japanese Soc. Artif. Intell. **14**(5), 771–780 (1999)
15. Golovin, D., Krause, A., Ray, D.: Near-optimal bayesian active learning with noisy observations. In: NIPS (2010)
16. Gretton, A., Borgwardt, K., Scholkopt, B.: A kernel method for the two-sample-problem. In: NIPS (2006)
17. Han, Y., Wu, F., Zhuang, Y., He, X.: Multi-label transfer learning with sparse representation. TCSVT **20**, 1110–1121 (2010)
18. Han, Y., Wu, F., Tao, D., Shao, J., Zhuang, Y., Jiang, J.: Sparse unsupervised dimensionality reduction for multiple view data. TCSVT **22**, 1485–1496 (2012)
19. Han, Y., Yang, Y., Ma, Z., Shen, H., Sebe, N., Zhou, X.: Image attribute adaptation. TMM **16**, 1115–1126 (2014)
20. Hoi, S., Jin, R., Lyu, M.: Large-scale text categorization by batch mode active learning. In: WWW (2006)
21. Hua, G., Long, C., Yang, M., Gao, Y.: Collaborative active learning of a kernel machine ensemble for recognition. In: ICCV (2013)
22. Huang, J., Smola, A., Scholkopf, B.: Correcting sample selection bias by unlabeled data. In: NIPS (2007)
23. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR (2011)
24. Li, X., Guo, Y.: Multi-level adaptive active learning for scene classification. In: ECCV (2014)
25. Liang, L., Grauman, K.: Beyond comparing image pairs: Setwise active learning for relative attributes. In: CVPR (2014)
26. Maron, O., Lozano-Perez, T.: A framework for multiple-instance learning. In: NIPS (1998)
27. Mihalkova, L., Huynh, T., Mooney, R.: Mapping and revising markov logic networks for transfer learning. In: AAAI (2007)
28. Natarajan, N., Dhillon, I.S., Ravikumar, P., Tewari, A.: Learning with noisy labels. In: NIPS (2013)
29. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowledge Data Eng. **22**(10), 1345–1359 (2010)
30. Rajagopal, A.K., Subramanian, R., Ricci, E., Vieriu, R.L., Lanz, O., Sebe, N., et al.: Exploring transfer learning approaches for head pose classification from multi-view surveillance images. IJCV **109**(1–2), 146–167 (2014)
31. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV (2010)
32. Sheng, V., Provost, F., Ipeirotis, P.: Get another label? Improving data quality and data mining using multiple, noisy labelers. In: KDD (2008)
33. Shi, X., Fan, W., Ren, J.: Actively transfer domain knowledge. In: ECML (2008)
34. Sogawa, Y., Ueno, T., Kawahara, Y., Washio, T.: Active learning for noisy oracle via density power divergence. Neural Netw. **46**, 133–143 (2013)

35. Stiefelhagen, R., Bowers, R., Fiscus, J.G.: Multimodal technologies for perception of humans. CLEAR, 2007 (2007)
36. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: ICML (2000)
37. Wang, X., Huang, T.K., Schneider, J.: Active transfer learning under model shift. In: ICML (2014)
38. Yan, R., Yang, J., Hauptmann, A.G.: Automatically labeling video data using multi-class active learning. In: ICCV (2003)
39. Yan, Y., Ricci, E., Subramanian, R., Lanz, O., Sebe, N.: No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In: ICCV (2013)
40. Yan, Y., Subramanian, R., Lanz, O., Sebe, N.: Active Transfer Learning for Multiview Head-pose Classification. ICPR (2012)
41. Yan, Y., Yang, Y., Shen, H., Meng, D., Liu, G., Hauptmann, A., Sebe, N.: Complex event detection via event oriented dictionary learning. In: AAAI (2015)
42. Yan, Y., Ricci, E., Subramanian, R., Liu, G., Sebe, N.: Multitask Linear Discriminant Analysis for View Invariant Action Recognition. IEEE Transactions on Image Processing, vol. 23, no. 12, (2014)
43. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: ACM MM (2007)
44. Yang, L., Hanneke, S., Carbonell, J.: A theory of transfer learning with application to actively transfer. JMLR (2012)
45. Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., Pan, Y.: A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. TPAMI **34**, 723–742 (2012)
46. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. IJCV, 11 (2014)
47. Yao, Y., Dorretto, G.: Boosting for transfer learning with multiple sources. In: CVPR (2010)
48. Zhang, J., Han, Y., Tang, J., Hu, Q., Jiang, J.: What can we learn about motion videos from still images? In: ACM MM (2014)