Instilling Social to Physical: Co-Regularized Heterogeneous Transfer Learning

Ying Wei[§], **Yin Zhu**[§], **Cane Wing-ki Leung**[†], **Yangqiu Song**[‡], and **Qiang Yang**[§]
[§]Hong Kong University of Science and Technology, Hong Kong
[†]Wisers Research, Hong Kong
[‡]West Virginia University, Morgantown, WV, USA
[§]{yweiad,qyang}@cse.ust.hk, [§] zhuyin.nju@gmail.com, [†] cane.leung@gmail.com, [‡]yangqiu.song@mail.wvu.edu

Abstract

Ubiquitous computing tasks, such as human activity recognition (HAR), are enabling a wide spectrum of applications, ranging from healthcare to environment monitoring. The success of a ubiquitous computing task relies on sufficient physical sensor data with groundtruth labels, which are always scarce due to the expensive annotating process. Meanwhile, social media platforms provide a lot of social or semantic context information. People share what they are doing and where they are frequently in the messages they post. This rich set of socially shared activities motivates us to transfer knowledge from social media to address the sparsity issue of labelled physical sensor data. In order to transfer the knowledge of social and semantic context, we propose a Co-Regularized Heterogeneous Transfer Learning (CoHTL) model, which builds a common semantic space derived from two heterogeneous domains. Our proposed method outperforms state-of-the-art methods on two ubiquitous computing tasks, namely human activity recognition and region function discovery.

Introduction

Recent years there have been extensive research efforts on ubiquitous computing, as various physical sensors such as GPS, accelerometers and Wi-Fi become widely and inexpensively available on smart devices. Ubiquitous computing tasks, such as Human Activity Recognition (HAR), are enabling a wide spectrum of applications, ranging from healthcare (Stone and Skubic 2015) to environment monitoring (Lane et al. 2010).

The success of a ubiquitous computing task relies on sufficient annotated or groundtruth data as in many other machine learning tasks. In general, obtaining the labelled data is expensive and tedious, while annotating raw sensor readings is particular challenging. Take HAR as an example. One possibility is to ask annotators to provide activity labels in real time as soon as sensor readings are being generated, which may be inconsiderate, and even impractical in many situations (imagine an annotator has to perform labelling while "jogging" or "driving"). Another possibility is to have annotators scan through raw sensor readings and manually label activities they performed post-hoc. This is also difficult, if not impossible, since sensor readings are hardly human readable and understandable. Meanwhile, people nowadays proactively share happenings about and around them, as well as their whereabouts on social media platforms such as Twitter. Such platforms thus provide a huge and rich semantic repository of activities that people are performing at different time and locations.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

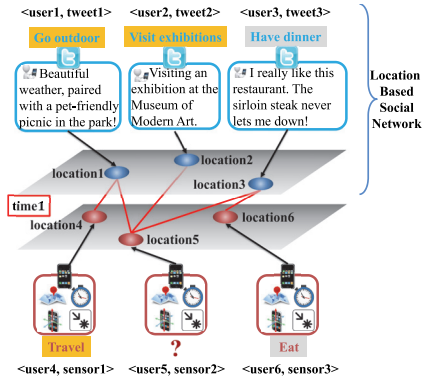


Figure 1: A motivating example on instilling knowledge from social media into sensors in the physical world.

Given that labelled sensor data are difficult to obtain, and that social media messages capture rich information about physical activities and locations, an interesting question arises: *can we transfer knowledge from social media side to the physical world to solve ubiquitous computing tasks*? We show in this paper that the answer is *yes*. On one hand, ubiquitous computing tasks have extensively exploited different physical sensors, including wearable sensors (Lara and Labrador 2013), GPS traces (Lin and Hsu 2014), Wi-Fi (Wang et al. 2014), barometer, temperature, and humidity sensors (Choudhury et al. 2008), etc. On the other hand, some recent studies on extracting human activities (Song et al. 2013) and events (Ritter et al. 2015) from social media have been reported. Yet, to the best of our knowledge, no previous work has considered integrating social media knowledge into physical sensor data. A seemingly related strand of research that combines the social and physical is Location Based Social Networks (LBSN). The studies in the field of LBSN either tackle link prediction or user profiling

on social side with the help of locations (Cho, Myers, and Leskovec 2011; Scellato and Mascolo 2011), or address location or activity recommendation on physical side with the help of social media (Bhargava et al. 2015; Gao et al. 2015; Zheng et al. 2010). Our problem targeting ubiquitous computing applications on physical side obviously differs from them. We transfer knowledge between two domains with a clear gap, i.e., natural language text in social media and raw sensor readings such as accelerometer in the physical world, while they deal with only a single domain in which locations act as a dimension of feature shown in Figure 1. Consequently, they are more straightforward than our problem in view of the data alignment because many social platforms have the data, i.e., messages and corresponding check-ins.

The challenges of bridging the gap between physical sensors and social media lie in both the mismatch of feature spaces and the missing of direct aligned data across the two domains. First, physical sensor features and social media messages have incommensurable representation structures. For example, a tri-axial accelerometer reading is represented as numerical values in three dimensions, whereas a tweet is oftentimes represented as a bag-of-words. Second, we have no access to directly aligned data out of privacy concerns which associate the same user's social media messages with his sensor records (e.g., via the user's smartphone). Such alignment doubtlessly facilitates the knowledge transfer.

Instead, we are only provided with two categories of prior knowledge that could indirectly align. One is partial label information which tell whether a sensor record and a message are semantically related. As shown in Figure 1, sensor1 carrying the label "Travel" is related to tweet1 and tweet2 which are labelled as "Go outdoor" and "Visit exhibitions", respectively. The label space of social media messages is obviously more diverse and detailed than that of sensor data, because natural language has more descriptive power than raw sensor data in nature. So this is also one of the concerns that we have to address. The other is the spatio-temporal information. We reasonably make a simplification: a sensor record and a social message which are spatio-temporally close enough share the same semantic meaning. The soundness of such simplification is guaranteed by our data which show that averagely 75% co-occurred sensor records and Weibo messages within a geographical distance (0.25 km) and a time period (1 hour) share the semantic meaning, e.g., activity for the HAR task. In Figure 1, we could infer user5's activity at time1 according to 1) the unlabelled sensor record sensor2, and 2) activities happening in his geographical proximity at the moment, which can be inferred from the tweets that Twitter users (user1, user2, user3) post.

In this paper, we address the two challenges, i.e., the mismatch of feature spaces and the missing of direct aligned data, by proposing a Co-Regularized Heterogeneous Transfer Learning (CoHTL) model. The model alleviates the feature space mismatch by projecting both domains onto a latent semantic subspace where they are comparable. The principles of extracting the subspace include: 1) the structure of instances in each domain is preserved; and 2) the semantic similarities between instances across the two domains are preserved. We formulate CoHTL as a matrix factorization model with co-regularization, in which matrix factorization maximizes the empirical likelihood to guarantee the first principle while co-regularization guards the second. Moreover, co-regularization with capability of incorporating both prior knowledge for alignment addresses the second challenge. As such, the feature representation of physical sensor data is enriched by social media messages in the subspace. We empirically show that the enriched feature representation is more discriminative, especially among activities that may generate similar raw sensor readings. The more discriminative power ensures that fewer labelled training sensor data are required, thereby addresses the sparsity of labelled data.

Instilling Knowledge from Social to Physical

Problem Formulation

Suppose that there are a few labelled instances of physical sensor data $\mathbf{P}_l = \{\mathbf{p}_i\}_{i=1}^{m_l}$ and some test sensor records $\mathbf{P}_t = \{\mathbf{p}_i\}_{i=m_l+1}^{m_l+m_t}$, where $\mathbf{p}_i = (\text{time}, \text{loc}, \mathbf{ac}) \in \mathbb{R}^p$ is the sensor feature vector. $\mathbf{p}_i$.time and $\mathbf{p}_i$.loc denote respectively the time and location $\mathbf{p}_i$ occurred, while $\mathbf{p}_i.\mathbf{ac}$ contains other physical sensor readings, such as accelerometer. $\mathbf{y}_l = \{y_i\}_{i=1}^{m_l}$ is the label vector associated with $\mathbf{P}_l$, while $\mathbf{y}_t = \{y_i\}_{i=m_l+1}^{m_l+m_t}$ corresponds to groundtruth labels of $\mathbf{P}_t$ for evaluation. Note that $y_i \in \{l_1^y, l_2^y, \cdots, l_a^y\}$, and that $\mathbf{P} = \mathbf{P}_l \cup \mathbf{P}_t \in \mathbb{R}^{m \times p}$ is the complete sensor data matrix where $m = m_l + m_t$. We are also provided with abundant message instances $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{n}$ on social media where $\mathbf{q}_j = (\text{time}, \text{loc}, \mathbf{wd}) \in \mathbb{R}^q$. $\mathbf{q}_j$.time and $\mathbf{q}_j$.loc indicate respectively the time and location attached to the message $\mathbf{q}_j$, and $\mathbf{q}_j.\mathbf{wd}$ is a bag-of-words feature vector representing the content of $\mathbf{q}_j$. We assume the existence of a function $g(\mathbf{q}_j) \to g_j$ to map each message $\mathbf{q}_j$ to $g_j \in \{l_1^g, l_2^g, \cdots, l_b^g\}$ as semantic supervision information. Our final goal is to learn a new representation of $\mathbf{p}_i$ in a u-dimensional latent semantic space, i.e., $\mathbf{u}_i \in \mathbb{R}^u$, so that we can classify activities better using $\mathbf{u}_i$ instead of original $\mathbf{p}_i$. We use boldface lowercase letters and uppercase letters to denote vectors and matrices, respectively. For a vector $\mathbf{x}$, $\|\mathbf{x}\|$ denotes its ℓ^2 norm. For a matrix $\mathbf{X}$, $\|\mathbf{X}\|_F^2$ denotes its Frobenius norm.

Bridging the Physical and the Social

We now illustrate how to bridge physical sensor records and messages on social media with a semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, in which $\mathbf{S}_{ij}$ indicates the extent to which the ith physical sensor record and the jth social message are semantically correlated.

Heterogeneous Label Alignment Our goal is to align a labelled sensor record $\mathbf{p}_i \in \mathbf{P}_l$ with a message $\mathbf{q}_j$ using their corresponding supervision information y_i and g_j. The label space of physical sensor data tends to differ from that of social messages. Using the labels of our HAR experiments as an example, Figure 2(a) lists all labels of our physical sensor data, while Figure 2(c) shows a sample of tags summarizing user activities in our social media data. Here, we propose two solutions to obtain the similarity s_{ab} between a physical sensor label l_a^y and a social message label l_b^g.

Topic model based: first of all, we construct each label in

Table 1: The first row shows the sensor labels for activity recognition, and the second and third row show the corresponding most semantically related social media labels determined by topic model based and word embedding based methods, respectively.

rest	drive	work	eat	exercise	travel
watch performances	drive halfway	work	have dinner	exercise	go outdoor
work	drive halfway	work	breakfast	exercise	buy tickets

each domain as a query which is then used to query an open knowledge base, such as World Wide Web, for auxiliary documents. We assume that the retrieved auxiliary documents are much relevant to each label. Since searching is not the focus of our work, we omit the details here. Secondly, each label is represented as the topic distribution vector of its auxiliary documents by performing Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). Thirdly, we adopt Jensen-Shannon divergence (Lin 1991), a symmetric metric to measure the similarity between two probability distributions, to measure the similarity s_{ab} between l_a^y and l_b^g in terms of their topic distribution vectors.

Word embedding based: we learn the embeddings for each label using word2vec (Mikolov et al. 2013), a neural network based language model that learns word embeddings. We train the skip-gram architecture of word2vec on the English Wikipedia corpus with context size equal to 5. Finally, we measure the similarity s_{ab} between l_a^y and l_b^g by computing the cosine similarity of their corresponding embeddings.

Consequently, the similarity s_{ij} between $\mathbf{p}_i \in \mathbf{P}_l$ and $\mathbf{q}_j$ equals to s_{ab} if $\mathbf{p}_i$'s label $y_i = l_a^y$ and $\mathbf{q}_j$'s label $g_j = l_b^g$. Results in Table 1 show that our heterogeneous label alignment methods are effective.

Spatio-temporal Alignment As detailed in the introduction, we make a simplification: if a physical sensor record and a social media message are spatio-temporally close to each other, they indicate the same activity. Motivated by this, we define the similarity s_{ij} between an unlabelled physical sensor record $\mathbf{p}_i \in \mathbf{P}_t$ and a message $\mathbf{q}_j$ as:

$$s_{ij} = \begin{cases} 1 & \text{if} \quad d_1^{ij} = 0 \quad \text{and} \quad d_2^{ij} \le r \\ b & \text{if} \quad d_1^{ij} \neq 0 \quad \text{and} \quad d_2^{ij} \le r \\ br/d_2^{ij} & \text{otherwise,} \end{cases}$$

where $d_1^{ij} = d_1(\mathbf{p}_i.\text{time}, \mathbf{q}_j.\text{time})$ evaluates the time distance and equals to 0 if and only if $\mathbf{p}_i$.time and $\mathbf{q}_j$.time are in the same hour of the same day of a week, e.g., both occurring during 3pm - 4pm on Wednesday. $d_2^{ij} = d_2(\mathbf{p}_i.\text{loc}, \mathbf{q}_j.\text{loc})$ is the geographical distance measured according to latitudes and longitudes. r is the geographical vicinity radius and $0 \le b \le 1$. We empirically determine $r = 0.25\text{km}, b = 0.9$ in the experiments.

Unifying Label and Spatio-temporal Similarity We denote the similarity matrix between $\mathbf{P}_l$ and $\mathbf{Q}$ obtained by heterogeneous label alignment as $\mathbf{S}_l \in \mathbb{R}^{m_l \times n}$, and the one between $\mathbf{P}_t$ and $\mathbf{Q}$ obtained by spatio-temporal alignment as $\mathbf{S}_t \in \mathbb{R}^{m_t \times n}$. Due to the inconsistent similarity metrics adopted by $\mathbf{S}_l$ and $\mathbf{S}_t$, we cannot simply combine them into the global similarity matrix $\mathbf{S}$. Instead, we adopt a local scheme, i.e., k-nearest-neighbour, to unify $\mathbf{S}_l$ and $\mathbf{S}_t$. In detail, for each row vector of either $\mathbf{S}_l$ or $\mathbf{S}_t$, we set the values of the k largest components, i.e., the k most semantically related messages to a sensor record, as "1" and the rest as "0". The value of k is empirically determined in our work.

Co-Regularized Heterogeneous Transfer Learning

Here, we discuss in detail how to find a latent semantic subspace onto which both domains are projected. The optimal subspace is defined as follows.

Definition 1 *Given the sensor data matrix $\mathbf{P} \in \mathbb{R}^{m \times p}$ and the message data matrix $\mathbf{Q} \in \mathbb{R}^{n \times q}$, the optimal projections of $\mathbf{P}$ and $\mathbf{Q}$ onto the optimal subspace, i.e., $\mathbf{U} \in \mathbb{R}^{m \times u}$ and $\mathbf{W} \in \mathbb{R}^{n \times u}$ respectively, are given by minimizing the following objective:*

$$\min_{\mathbf{U},\mathbf{W}} \ell(\mathbf{P}, \mathbf{U}) + \ell(\mathbf{Q}, \mathbf{W}) + \beta \mathbf{D}(\mathbf{U}, \mathbf{W}), \qquad (1)$$

where $\ell(\cdot, \cdot)$ is a distortion function that evaluates the difference between the original data and projected data (e.g., $\mathbf{P}$ and $\mathbf{U}$). $\mathbf{D}(\cdot, \cdot)$ is the co-regularizer which encourages the pairwise similarities between projected data of the two domains (i.e., between $\mathbf{U}$ and $\mathbf{W}$) to be consistent with the original semantic similarities. β is a trade-off parameter.

On the one hand, according to the first two terms in Equation (1), we expect the projections to preserve the structures of the original data as much as possible. We achieve this goal by defining $\ell(\cdot, \cdot)$ as the following,

$$\ell(\mathbf{P}, \mathbf{U}) + \ell(\mathbf{Q}, \mathbf{W}) = \|\mathbf{P} - \mathbf{U}\mathbf{V}_1\|_F^2 + \|\mathbf{Q} - \mathbf{W}\mathbf{V}_2\|_F^2, \quad (2)$$

in which we factorize the original data into the projections ($\mathbf{U}$ and $\mathbf{W}$) and the linear mapping matrices ($\mathbf{V}_1$ and $\mathbf{V}_2$). Note that Matrix Factorization is widely known as an effective tool to extract latent subspaces while preserving the original data's structures by maximizing the empirical likelihood. In a different light, $\mathbf{V}_1^T \in \mathbb{R}^{p \times u}$ and $\mathbf{V}_2^T \in \mathbb{R}^{q \times u}$ map $\mathbf{P}$ and $\mathbf{Q}$, respectively, into a u-dimensional space where the projected data are comparable.

On the other hand, Equation (1) introduces a co-regularizer to ensure that the optimal projections $\mathbf{U}$ and $\mathbf{W}$ in the latent space should preserve the semantic similarities between the original physical sensor records and messages. Thus, we define the co-regularizer as:

$$\mathbf{D}(\mathbf{U}, \mathbf{W}) = \sum_{i=1}^{m} \sum_{j=1}^{n} S_{ij} \|\mathbf{u}_i - \mathbf{w}_j\|_2^2, \qquad (3)$$

where $S_{ij} = (\mathbf{S})_{ij}$. $\mathbf{S}$ is the similarity matrix we have obtained. Minimizing this term enforces two semantically similar examples' projections to be as close as possible in the latent subspace, otherwise incurring a heavy penalty.

The trade-off parameter β balances the importance of original structure preservation and semantic similarity co-regularization. Furthermore, for each sensor record, the similarity matrix $\mathbf{S}$ delicately selects the best social media messages to transfer. The larger the similarity, the more likely the

corresponding social media messages we transfer. To avoid "negative transfer" (Pan and Yang 2010), we only select a subset of messages on social media, which geographically overlap with our target physical sensor records, e.g., in a city. We would believe that such data are always available in view of massive social media datasets nowadays.

Optimization

Substituting Equations (2) and (3) into Equation (1), we obtain the objective function to minimize *w.r.t.* $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}_1$ and $\mathbf{V}_2$ as follows:

$$\mathcal{O} = \|\mathbf{P} - \mathbf{U}\mathbf{V}_1\|_F^2 + \|\mathbf{Q} - \mathbf{W}\mathbf{V}_2\|_F^2 + \beta \sum_{i=1}^{m}\sum_{j=1}^{n} S_{ij}\|\mathbf{u}_i - \mathbf{w}_j\|_2^2 + \gamma R(\mathbf{U}, \mathbf{W}, \mathbf{V}_1, \mathbf{V}_2), \quad (4)$$

where $R(\mathbf{U}, \mathbf{W}, \mathbf{V}_1, \mathbf{V}_2) = \|\mathbf{U}\|_F^2 + \|\mathbf{W}\|_F^2 + \|\mathbf{V}_1\|_F^2 + \|\mathbf{V}_2\|_F^2$ is the regularization term which controls the complexity of $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}_1$, $\mathbf{V}_2$. The optimization problem in Equation (4) is not jointly convex *w.r.t.* the four matrices $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}_1$ and $\mathbf{V}_2$, thus only has local optimal solutions. However, it is convex *w.r.t.* any one of them while fixing the other three. We therefore adopt an alternating algorithm to solve this problem, by iteratively fixing three of the matrices to solve the remaining one until convergence. We define our algorithm formally as follows.

Fix $\mathbf{U}$, $\mathbf{V}_1$, $\mathbf{V}_2$: in this case, the objective *w.r.t.* $\mathbf{W}$ is:

$$\mathcal{O}_w = Tr[(\mathbf{Q} - \mathbf{W}\mathbf{V}_2)(\mathbf{Q} - \mathbf{W}\mathbf{V}_2)^T] + \beta Tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W}) - 2\beta Tr(\mathbf{U}^T\mathbf{S}\mathbf{W}) + \gamma Tr(\mathbf{W}\mathbf{W}^T), \quad (5)$$

where we introduce a diagonal matrix $\mathbf{S}_w$ with diagonal elements $(\mathbf{S}_w)_{jj} = \sum_{i=1}^{m} S_{ij}$. Equation (5) is convex *w.r.t.* $\mathbf{W}$, so we can use any gradient-based method to find the local minimum. Here we adopt the conjugate gradient descent method with the gradient provided as below:

$$\frac{\partial \mathcal{O}_w}{\partial \mathbf{W}} = 2[-\mathbf{Q}\mathbf{V}_2^T + \mathbf{W}\mathbf{V}_2\mathbf{V}_2^T + \beta\mathbf{S}_w\mathbf{W} - \beta\mathbf{S}^T\mathbf{U} + \gamma\mathbf{W}]. \quad (6)$$

Fix $\mathbf{W}$, $\mathbf{V}_1$, $\mathbf{V}_2$: similarly, the objective *w.r.t.* $\mathbf{U}$ is:

$$\mathcal{O}_u = Tr[(\mathbf{P} - \mathbf{U}\mathbf{V}_1)(\mathbf{P} - \mathbf{U}\mathbf{V}_1)^T] + \beta Tr(\mathbf{U}^T\mathbf{S}_u\mathbf{U}) - 2\beta Tr(\mathbf{U}^T\mathbf{S}\mathbf{W}) + \gamma Tr(\mathbf{U}\mathbf{U}^T), \quad (7)$$

where $\mathbf{S}_u$ is also diagonal with diagonal elements $(\mathbf{S}_u)_{ii} = \sum_{j=1}^{n} S_{ij}$. The corresponding gradient is given by:

$$\frac{\partial \mathcal{O}_u}{\partial \mathbf{U}} = 2[-\mathbf{P}\mathbf{V}_1^T + \mathbf{U}\mathbf{V}_1\mathbf{V}_1^T + \beta\mathbf{S}_u\mathbf{U} - \beta\mathbf{S}\mathbf{W} + \gamma\mathbf{U}]. \quad (8)$$

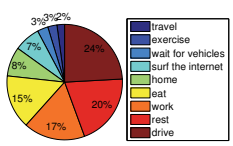
Fix $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}_2$: we employ the multivariate ridge regression model (Hoerl and Kennard 1970) to update $\mathbf{V}_1$ and yield the following solution:

$$\mathbf{V}_1 = (\mathbf{U}^T\mathbf{U} + \gamma\mathbf{I}_{u\times u})^{-1}\mathbf{U}^T\mathbf{P}. \quad (9)$$

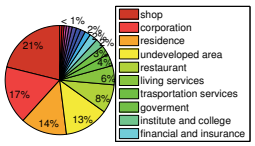
Fix $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}_1$: similar to $\mathbf{V}_1$, we give the update of $\mathbf{V}_2$:

$$\mathbf{V}_2 = (\mathbf{W}^T\mathbf{W} + \gamma\mathbf{I}_{u\times u})^{-1}\mathbf{W}^T\mathbf{Q}. \quad (10)$$

Here we briefly analyze the time complexity of our algorithm. The first computationally expensive part is to evaluate the gradients during optimizing $\mathbf{W}$ and $\mathbf{U}$. The time complexity is $O((nq + mp + N)u)$ where N counts the number of non-zero elements in $\mathbf{S}$. According to the construction of $\mathbf{S}$, we know $N \ll mn$. Another major computation cost comes from updating $\mathbf{V}_1$ and $\mathbf{V}_2$. The time complexity is $O(u^2(m+n)+u^3+u(mp+nq))$. Since the dimension of the subspace, i.e., u, is small, u^3 and u^2 are not that computationally expensive. We conclude that our algorithm scales linearly with the number of physical sensor records (m), the number of social messages (n), the number of non-zero elements in $\mathbf{S}$ (N), and the data dimensionality (p and q).



(a) The category distribution of activities in the physical sensor dataset.



(b) The category distribution of functionalities in the trajectories featured region dataset. Only top ten functionalities are in legend.

take beauty or massage	home	stay in a hotel	go outdoor	go to school	take public transport
exercise	have breakfast	have drinks	visit exhibitions	work	...

(c) 11 out of 30 selected labels extracted in Sina Weibo dataset.

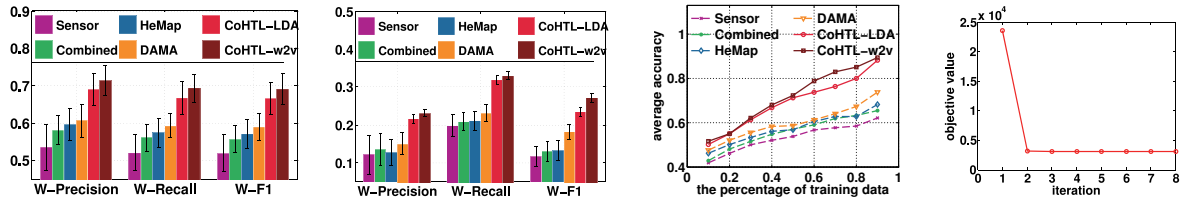
Figure 2: Overview of the datasets.

Experiments

Datasets

We verify the effectiveness of social knowledge transfer on two ubiquitous computing tasks, namely human activity recognition and region function discovery. In the activity recognition task, we collected 232 sensor records through cellphones from 10 volunteers. Each sensor record contains time, GPS, tri-axial accelerometer, and POI information. Figure 2(a) shows the distribution of activity labels of the 232 sensor records. In the region function discovery task, our goal is to discover the functionality of a region (e.g., "residence"). Our dataset is a collection of taxis trajectories generated by 182 taxis within one month in a big city of South China whose population is over 10 million. There exist 6,010 regions, each of which is a $0.25\times0.25\text{km}^2$ grid. For each region, we extract in-region and out-region taxis counts, average taxis speed, and duration of stay in each hour as features. Figure 2(b) shows the distribution of groundtruth functionalities of all the regions.

We obtain social knowledge from Sina Weibo, a Twitter-like microblogging service in China. The full dataset contains tweets from about 10 million users. We use all 10,791 tweets associated with check-ins that lie in a geographical bounding box determined by all physical sensor records' GPS locations. We obtain the activity label of a tweet using the rule-based algorithm in (Song et al. 2013), making use of a tweet's location, posting time and named entities in the text. Figure 2(c) selects some of the activity labels to present.



(a) Performance comparison on activity recognition.

(b) Performance comparison on region function discovery.

(c) Varying the percentage of training sensor records.

(d) Convergence analysis.

Figure 3: Performance comparison.

Evaluation Metrics and Baselines

We use weighted precision (W-Precision), weighted recall (W-Recall), weighted F1 score (W-F1), and accuracy as our evaluation metrics. Weighted F1 score is computed by summing weighted F1 scores of all categories. The weight of a category's F1 score is the percentage of instances in the category. Weighted precision and recall have similar definitions.

Our models, using topic model based and word embedding based methods for heterogeneous label alignment, are denoted as **CoHTL-LDA** and **CoHTL-w2v**, respectively. We compare them with the following four baselines:

Sensor. This method only uses original physical sensor features without social knowledge transfer. The majority of recent work on activity recognition and region function discovery fall into this category, e.g., (Lara and Labrador 2013).

Combined. We directly concatenate features from both the physical and social. But physical sensor records $\mathbf{P} \in \mathbb{R}^{m \times p}$ and social messages $\mathbf{Q} \in \mathbb{R}^{n \times q}$ are not directly aligned. Therefore we obtain the aligned social messages matrix $\widetilde{\mathbf{Q}} \in \mathbb{R}^{m \times q}$ first: for each physical sensor record, we sum up the features of all tweets within a certain spatio-temporal threshold to this sensor. The baseline directly combines $\mathbf{P}$ and $\widetilde{\mathbf{Q}}$ as the final feature matrix which is normalized.

HeMap. Heterogeneous Spectral Mapping (HeMap) (Shi et al. 2013) projects data in two domains with correspondence onto a latent space. We adapt HeMap into our problem by taking $\mathbf{P}$ and $\widetilde{\mathbf{Q}}$ as its input. HeMap, however, cannot incorporate any label information for alignment.

DAMA. Heterogeneous Domain Adaptation with Manifold Alignment (DAMA) (Wang and Mahadevan 2011) aligns different domains into a latent space using label information based manifold regularization. However, DAMA only works on the data that have strong manifold structures, and does not handle heterogeneous label spaces for different domains. We adapt DAMA into our problem by letting our similarity matrix $\mathbf{S}$ (in which the label alignment is by word embedding based method) to substitute its label alignment matrix.

Upon different feature representations obtained by different models, we use linear SVM (Chang and Lin 2011) as the base classifier. The trade-off parameter C of linear SVM is set according to 10-fold cross validation for each model. We repeat the experiments 30 times and report average results.

Activity Recognition

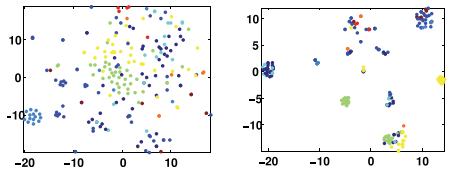
Performance comparison In this experiment, we randomly sample 100 sensor records as training samples to perform 9-class classification, and the other 132 sensor records as test data. Figure 3(a) shows the comparison. Our methods show significant, up to 20%, improvement over other methods. A naive combination of sensor and social features performs better than sensor features only (Combined v.s. Sensor), which validates the necessity of instilling social knowledge into physical sensor data. HeMap shows a little improvement over Combined, because employing social messages to enrich sensor readings' feature representation in a latent space is more effective than naive combination. DAMA outperforms HeMap because DAMA is adapted to regularize with $\mathbf{S}$ which incorporates heterogeneous label alignment besides indirect correspondence. To preserve the structure of instances in each domain, DAMA assumes that the data in each domain lie in a manifold, while our method CoHTL naturally maximizes the empirical likelihood without any assumption, thus defeats DAMA. Moreover, our model with word embedding based method for heterogeneous label alignment shows more superiority than with topic model based method. In Figure 4, we compare the confusion matrix obtained by our method with the one by Sensor. In general, our method improves the recalls and precisions of all categories a lot. For activities that may generate similar sensor readings, such as "work" and "rest", Sensor cannot discriminate between them. Our method however presents both higher precisions and recalls when classifying such activities. Note that in this experimental setting our algorithm converges within 10 iterations as Figure 3(d) shows.

Quality of feature representation In Figure 5(a) and 5(b), we examine the quality of feature representation in the subspace by employing t-SNE (Van der Maaten and Hinton 2008) to visualize the 2D projection. Due to space limitation, we only compare CoHTL-w2v with DAMA. The experiment setting is the same as in Figure 3(a). Instances with the same label, in the same color, show more obvious clustering structures in the latent space obtained by our method, which explains the superior classification performance of CoHTL.

Varying the number of training sensor records Figure 3(c) shows that our methods can handle extremely sparse training data and greatly improve classification accuracies when more training data are available. CoHTL-w2v improves almost 25% over Sensor when only 10% training examples are available. As the number of training examples increases,

	rest	drive	home	surf the internet	work	eat	exercise	travel	wait vehicles	recall
					Sensor					
rest	13	3	0	5	0	2	2	0	0	0.52
drive	3	17	0	0	0	2	0	0	1	0.73913
home	0	3	9	0	0	3	0	0	0	0.6
surf the internet	3	2	0	2	0	0	0	0	0	0.28571
work	6	2	1	1	22	2	0	1	0	0.62857
eat	1	2	0	1	0	9	0	0	1	0.64285
exercise	0	1	0	0	0	0	1	1	0	0.33333
travel	0	1	0	0	0	0	0	0	0	0
wait vehicles	0	0	0	0	0	1	0	0	1	0.5
precision	0.5	0.54838	0.9	0.22222	1	0.47368	0.33333	0	0.33333	
					CoHTL-w2v					
rest	18	1	0	6	0	4	1	0	0	0.6
drive	2	22	0	1	0	3	1	0	2	0.70967
home	0	2	9	0	0	0	0	0	0	0.81818
surf the internet	2	1	0	2	0	1	0	1	0	0.28571
work	2	1	1	0	22	1	0	0	0	0.81481
eat	2	2	0	0	0	10	0	0	0	0.71428
exercise	0	0	0	0	0	0	1	0	0	1
travel	0	1	0	0	0	0	0	1	0	0.5
wait vehicles	0	1	0	0	0	0	0	0	1	0.5
precision	0.692308	0.70967	0.9	0.22222	1	0.52631	0.33333	0.5	0.33333	

Figure 4: Confusion matrix comparison of Sensor and CoHTL.



(a) DAMA (b) CoHTL-w2v

Figure 5: t-SNE visualization of the latent feature space.

our model dramatically improves as it is capable of taking advantage of labels for alignment.

Parameter sensitivity We also study the effects of two parameter settings, i.e., β and u, on the performance of CoHTL. The average accuracy of 10-fold cross validation on training data is examined. We perform grid search on β by fixing u. CoHTL gains the best average accuracy at $\beta = 0.1$ as Figure 6(a) shows. We understand that over-regularization of semantic similarities across two domains deteriorates the classification performance, probably because the original physical sensor data's structure has been changed. When fixing β, grid search of u shows that $u = 300$ performs the best. We adopt $\beta = 0.1, u = 300$ in our experiments.

Region Function Discovery

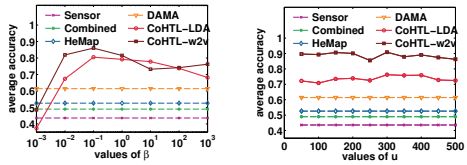
We only present the performance comparison of different methods on this task due to space limitation. Here we consider an extremely sparse case, where we take only 1% of all 6,010 regions as training instances to train a 20-class classifier. Figure 3(b) shows that our methods outperform the baselines by almost 50% in terms of weighted F1 scores. Thus, when there are few training data, our model can fully take advantage of social knowledge by aligning two domains with partial labels as well as the indirect correspondence.

Related Work

We outline related work under two topics, namely human activity recognition and heterogeneous transfer learning.

Human Activity Recognition

Human activity recognition (HAR) has been an increasingly popular research field, along the way from the earliest wearable inertial sensors (Bao and Intille 2004) to mobile



(a) Varying values of β (b) Varying values of u

Figure 6: Study of parameter settings.

sensing (Lane et al. 2010). Unfortunately, HAR still suffers from the lack of annotated grountruth data. Some studies (Cook, Feuz, and Krishnan 2013; Hu and Yang 2011; Zheng, Hu, and Yang 2009) tried to alleviate this problem by transferring labelled data from one activity recognition task to another, but such studies were limited to transferring between physical sensor data only.

In the social computing area, some studies on detecting events (Ritter et al. 2015) and classifying user activities (Song et al. 2013) via mining millions of social media messages have been reported. These studies focused on extracting activities implied or mentioned in text.

The most seemingly related work (Bhargava et al. 2015; Sattari et al. 2012; Zheng et al. 2010) make location or activity recommendation in Location Based Social Networks (LBSN). The authors used tags or tips that users post in a check-in location to represent this location's activity, so that users' interests can be modelled for recommendation of locations or activities by examining users' check-in histories. (Bhargava et al. 2015) also investigated the influence of user-user social relationship on users' choices of locations and activities. However, our work obviously tackles a different problem, i.e., activity recognition instead of recommendation. Besides, as we mentioned in the introduction, we transfer knowledge across two domains while these three work actually focused on a single domain in which the locations and social posts are well-matched features. As a result, they are more straightforward since many LBSN platforms nowadays provide social posts and associated check-ins.

Heterogeneous Transfer Learning

(Yang et al. 2009) first proposed heterogeneous transfer learning which transfers knowledge across domains in different feature spaces. While this work tackled the clustering task, later a series of studies (Dai et al. 2008; Duan, Xu, and Tsang 2012; Li et al. 2014; Shi et al. 2013; Wang and Mahadevan 2011; Zhu et al. 2011) focused on classification. To bridge the gap between different domains, the models including TLRisk (Dai et al. 2008), HTLIC (Zhu et al. 2011) that follows Collective Matrix Factorization (Singh and Gordon 2008), and HeMap (Shi et al. 2013), make use of direct correspondence data, e.g., tagged images that bridge images and text. The other group of models including DAMA (Wang and Mahadevan 2011), HFA (Duan, Xu, and Tsang 2012), and SHFA (Li et al. 2014), assumes the availability of abundant labelled data and aligns different domains with labels. However, our model, with only spatio-temporal information and partial labels provided, has to harness the collective power of these two prior knowledge. Besides, no

existing work handles the situation where label spaces of two domains are different. Most importantly, these methods are all used for applications involving images and text which are quite different from the ubiquitous computing applications.

Conclusion

In this paper, we propose a novel co-regularized heterogeneous transfer learning model to improve the performance of ubiquitous computing tasks by transferring knowledge from messages on social media. The social knowledge enriches the feature representation of physical sensors, thereby addresses the sparsity of labelled data in such tasks and the ambiguity of sensor data. Extensive experimental results demonstrate the superiority of our proposed method. In the future, we consider a more sophisticated case where activities of spatio-temporally close physical sensor records and social messages follow a distribution instead of being the same, and expect further performance improvement.

Acknowledgements

We thank the reviewers for their valuable comments to improve this paper. The research has been supported by Hong Kong CERG research projects 16211214 and 16209715.

References

Bao, L., and Intille, S. S. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive computing*. 1–17.

Bhargava, P.; Phan, T.; Zhou, J.; and Lee, J. 2015. Who, What, When, and Where: Multi-Dimensional Collaborative Recommendations Using Tensor Factorization on Sparse User-Generated Data. In *WWW*, 130–140.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *TIST* 2(3):27.

Cho, E.; Myers, S.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *SIGKDD*, 1082–1090.

Choudhury, T.; Consolvo, S.; Harrison, B.; Hightower, J.; LaMarca, A.; LeGrand, L.; Rahimi, A.; Rea, A.; Bordello, G.; Hemingway, B.; et al. 2008. The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, IEEE* 7(2):32–41.

Cook, D.; Feuz, K. D.; and Krishnan, N. C. 2013. Transfer learning for activity recognition: A survey. *Knowledge and information systems* 36(3):537–556.

Dai, W.; Chen, Y.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *NIPS*, 353–360.

Duan, L.; Xu, D.; and Tsang, I. W. 2012. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 711–718.

Gao, H.; Tang, J.; Hu, X.; and Liu, H. 2015. Content-aware point of interest recommendation on location-based social networks. In *AAAI*.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1).

Hu, D. H., and Yang, Q. 2011. Transfer learning for activity recognition via sensor mapping. In *IJCAI*, volume 22, 1962.

Lane, N. D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T.; and Campbell, A. T. 2010. A survey of mobile phone sensing. *Communications Magazine, IEEE* 48(9):140–150.

Lara, O. D., and Labrador, M. A. 2013. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE* 15(3):1192–1209.

Li, W.; Duan, L.; Xu, D.; and Tsang, I. W. 2014. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *PAMI* 36(6):1134–1148.

Lin, M., and Hsu, W.-J. 2014. Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing* 12:1–16.

Lin, J. 1991. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on* 37(1):145–151.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *TKDE* 22(10):1345–1359.

Ritter, A.; Wright, E.; Casey, W.; and Mitchell, T. 2015. Weakly Supervised Extraction of Computer Security Events from Twitter. In *WWW*, 896–905.

Sattari, M.; Manguoglu, M.; Toroslu, I. H.; Symeonidis, P.; Senkul, P.; and Manolopoulos, Y. 2012. Geo-activity recommendations by using improved feature combination. In *UbiComp*, 996–1003.

Scellato, S., and Mascolo, C. 2011. Exploiting Place Features in Link Prediction on Location-based Social Networks Categories and Subject Descriptors. In *SIGKDD*, 1046–1054.

Shi, X.; Liu, Q.; Fan, W.; and Yu, P. S. 2013. Transfer across completely different feature spaces via spectral embedding. *TKDE* 25(4):906–918.

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *SIGKDD*, 650–658.

Song, Y.; Lu, Z.; Leung, C. W.-k.; and Yang, Q. 2013. Collaborative boosting for activity classification in microblogs. In *SIGKDD*, 482–490.

Stone, E. E., and Skubic, M. 2015. Fall detection in homes of older adults using the microsoft kinect. *Biomedical and Health Informatics, IEEE Journal of* 19(1):290–301.

Van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR* 9(2579-2605):85.

Wang, C., and Mahadevan, S. 2011. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 1541.

Wang, Y.; Liu, J.; Chen, Y.; Gruteser, M.; Yang, J.; and Liu, H. 2014. E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures. In *MobiCom*, 617–628.

Yang, Q.; Chen, Y.; Xue, G.-R.; Dai, W.; and Yu, Y. 2009. Heterogeneous transfer learning for image clustering via the social web. In *ACL*, 1–9.

Zheng, V. W.; Zheng, Y.; Xie, X.; and Yang, Q. 2010. Collaborative location and activity recommendations with gps history data. In *WWW*, 1029–1038.

Zheng, V. W.; Hu, D. H.; and Yang, Q. 2009. Cross-domain activity recognition. In *UbiComp*, 61–70.

Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S. J.; Xue, G.-R.; Yu, Y.; and Yang, Q. 2011. Heterogeneous transfer learning for image classification. In *AAAI*.