Transfer Learning for Activity Recognition in Mobile Health

Yuchao Ma∗
Amazon.com
Seattle, WA

Andrew T. Campbell
Dartmouth College
Hanover, NH

Diane J. Cook
Washington State University
Pullman, WA

John Lach
George Washington University
Washington, DC

Shwetak N. Patel
University of Washington
Seattle, WA

Thomas Ploetz
Georgia Institute of Technology
Atlanta, GA

Majid Sarrafzadeh
University of California Los Angeles
Los Angeles, CA

Donna Spruijt-Metz
University of Southern California
Los Angeles, CA

Hassan Ghasemzadeh†
Washington State University
Pullman, WA

ABSTRACT

While activity recognition from inertial sensors holds potential for mobile health, differences in sensing platforms and user movement patterns cause performance degradation. Aiming to address these challenges, we propose a transfer learning framework, TransFall, for sensor-based activity recognition. TransFall’s design contains a two-tier data transformation, a label estimation layer, and a model generation layer to recognize activities for the new scenario. We validate TransFall analytically and empirically.

CCS CONCEPTS
-Computing methodologies → Learning paradigms; -Human-centered computing → Ubiquitous and mobile computing systems;

KEYWORDS
activity recognition, mobile health, transfer learning

1 INTRODUCTION

The rapid integration of sensors technologies, fueled by computational algorithms, creates a unique opportunity for remote health monitoring, long-term fitness tracking, and fall detection. A core task to support these applications is activity recognition. However, differences in sensing platforms and user behavior has limited generalizability of the activity recognition models. For example, if a user wants to maintain the usability of a well-trained motion analysis app installed on the old device, the user would like to avoid providing additional manual annotations for model re-training on the new device.

![Figure 1: Sensor readings from two smartphone models.](image)

Activity recognition performance is also adversely impacted by sensor biases from low-quality modules and sampling rate instability [7]. For example, F1-score declines 34.4% when training and test data belong to different devices (e.g., Samsung Galaxy S3 vs LG Nexus). Figure 1 shows acceleration for one subject’s walking behavior gathered by two smartphones. Such data divergence can propagate through the data processing pipeline, leading to accuracy decline.

We propose TransFall to overcome the challenges caused by cross-domain variations while reducing the dependency on labeled training data. Shown in Figure 2, the framework starts with a two-tier data transformation layer based on marginal distribution matching approaches, followed by a label estimation layer using a kernel method encoded in a weighted least-mean-squares fitting, and ends with a model generation layer using the previous step’s label set.

Development of transfer learning algorithms for activity recognition is also central in the long-term goal of creating a person’s digital twin. A digital twin is a digital replica of the human subject, built from multiple information sources including mobile sensor data. This quantified self provides a platform to understand the

---

1 Work completed while at Washington State University.
2 Corresponding author (hassan.ghasemzadeh@wsu.edu).
3 TransFall code and sample data are available at https://github.com/y-max/TransFall.
relationship between human behavior and influencing factors including health, genetics, and the environment. A digital twin offers the basis for automating health assessment and evaluating potential interventions on digital prototypes. Fusing data from multiple sensing platforms also creates a set of unlabeled data samples, which can be drawn from a normal distribution with mean \( \mu_t \) and variance \( \sigma_t^2 \), denoted as \( x_t \sim N(\mu_t, \sigma_t^2) \).

To minimize the difference of the means and variances between column variables in \( X_s \) and \( X_t \), we perform a linear transformation on each dimension \( i \in [1, d] \) in \( X_t \) with respect to \( X_s \). The output of the vertical transformation is the transformed target dataset \( \tilde{X}_t \), where \( x_i^t \sim N(\mu_i, \sigma_i^2) \) for each \( i \in [1, d] \).

2.3 Horizontal Transformation

The horizontal transformation layer further reduces the discrepancies in multi-variate variables between \( X_s \) and \( \tilde{X}_t \) using importance sampling, a common method to address covariate shift [6]. This technique finds a weight factor \( \beta \) for \( X_s \) that assigns higher weight to those source data samples that are more representative of the target dataset. The goal of this transformation is as follows:

\[
\text{Minimize} \int_x |\beta(x)P_s(x) - P_t(x)|dx .
\]

Because distributions \( P_s(x) \) and \( P_t(x) \) are unknown in practice, we use a kernel-based algorithm, empirical Kernel Mean Matching (eKMM) [5], to find the optimal weight factor \( \beta \) with the use of Reproducing Kernel Hilbert Space (RKHS) technique. Let \( \Phi : X \mapsto \mathcal{F} \) be a function that maps a vector variable \( X \) onto a feature space \( \mathcal{F} \). The output of the eKMM algorithm is the optimal weight factor \( \beta \), which can minimize the distance between the empirical means of \( X_s \) and \( \tilde{X}_t \) on the feature space \( \mathcal{F} \), as shown in the following equations.

\[
\text{Minimize} \quad \frac{1}{N_t} \sum_{i=1}^{N_t} \beta_i \| \Phi(x_i^t) - \frac{1}{N_s} \sum_{j=1}^{N_s} \Phi(x_j^s) \|^2
\]

Subject to \( \beta_i \geq 0, \quad i \in [1, N_t] \), and \( \frac{1}{N_t} \sum_{i=1}^{N_t} \beta_i - 1 \leq \epsilon \).

Each feature \( f_j \in \mathcal{F} \) is computed independently. Hence, we use the naive Bayes approximation to factorize the joint probability distribution, or \( P(x) = P(x_1, \cdots, x_d) = \prod_{i=1}^d P(x_i) \).

Therefore, the optimization objective of TransFall can be viewed as the summation of \( d \) subordinate optimization problems for \( d \) column variables, where \( \phi_i \in \Phi \) is a mapping function that converts the original probability distribution of a one-dimensional variable into a different distribution.

\[
\text{Minimize} \sum_{i=1}^{d} \int x_i |P_s(x_i) - \phi_i(P_t(x_i))|dx_i
\]
Proposition 2.1. Given a positive definite kernel $k$ over a vector space $\mathcal{X}$, we can find a Hilbert space $\mathcal{H}$ and a mapping function $\Phi : \mathcal{X} \mapsto \mathcal{H}$, such that $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}$, where $x_i, x_j \in \mathcal{X}$.

With the use of a kernel function that is positive definite on Euclidean space $\mathbb{R}^d$, optimization (3) can be solved without explicitly defining the mapping function $\Phi$. For this purpose, we use Gaussian kernel. Therefore, the objective function (3) can be rephrased as:

$$
\text{Minimize } \frac{1}{2} \beta^T K \beta - \kappa^T \beta \quad (4)
$$

where the kernel matrix $K$ and the kernel expansion $\kappa$ are given by

$$
K_{ij} := k(x_i^t, x_j^t); \quad \kappa_i := \frac{N_t}{N_s} \sum_{j=1}^{N_t} k(x_i^t, x_j^t) \quad (5)
$$

As a result, the optimal $\beta$ can be determined by solving (4) with the constraints listed in (3) using quadratic programming. Note that, similar to the vertical transformation module, the horizontal transformation module has also the potential to be coupled with existing machine learning algorithms that support the sample re-weighting.

2.4 Label Estimation

Given the transformed target dataset $\tilde{X}_t$ and the weight factor $\beta$, which approximates the distribution of $\tilde{X}_t$ using the source dataset $X_s$, the label estimation module intends to estimate the label set $\tilde{Y}_t$ for $\tilde{X}_t$ in preparation for training an activity recognition model for the target domain. The label estimation objective can be written as follows, for $x_i \in X_s$:

$$
\text{Minimize } \sum_{i=1}^{N_s} [\tilde{y}_i - \beta(x_i)f(x_i)]^2
$$

We can rewrite this optimization problem using a weighted least-mean-squares (LMS) fitting technique with a 2-norm regularization term as shown below. The LMS technique is commonly used for parameter estimation in linear models [5],

$$
\text{Minimize } \sum_{i=1}^{N_s} \beta_i (y_i^t - f(x_i^t))^2 + \lambda \| f \|^2 \quad (6)
$$

However, the optimal function $f$ in (6) is not necessarily linear. Therefore, we convert (6) using a linear model, based on the representer theorem [2], to the following:

$$
\text{Minimize } \alpha \sum_{i=1}^{N_s} \beta_i (y_i^t - \sum_{j=1}^{N_t} \alpha_j k(x_j^t, x_i^t))^2 + \lambda \| \sum_{j=1}^{N_t} \alpha_j k(x_j^t, \cdot) \|^2
$$

which can be written (after extension) as shown in (7).

$$
\text{Minimize } (Y_t - Ka)^T \beta (Y_t - Ka) + \lambda \alpha^T \beta \quad (7)
$$

where $K$ represents the kernel matrix in (5) and $\beta$ is a $N_s \times N_t$ diagonal matrix of $\beta$. If $K$ and $\beta$ are full rank matrices, the optimal solution for $\alpha$ can be derived using:

$$
\alpha = (\lambda \beta^{-1} + K)^{-1} Y_t \quad (8)
$$

Because the label set in activity recognition often contains multiple activity classes, we use a one-to-all approach for label estimation, by first solving $L$ optimal linear models $a^m$ for all activity labels, and then combining $L$ corresponding estimations of the data sample $x_i^t \in \tilde{X}_t$, to make the final prediction.

$$
\tilde{y}_i^t = \text{argmax}_m \sum_{j=1}^{N_t} a^m_j k(x_j^t, x_i^t) \quad (9)
$$

3 EXPERIMENTAL RESULTS

We conducted experiments on three publicly-available datasets [1, 7], as listed in Table 1. We evaluated the performance of TransFall in three scenarios including cross-platform, cross-subject, and hybrid. We use the notations shown in Table 1 to refer to various datasets, transfer learning scenarios, and comparison approaches. The comparison approach IWLSPC refers to the importance-weighted least-squares probabilistic classifier (IWLSPC) approach introduced in [4]. IWLSPC combines a least-squares probabilistic model with a sample re-weighting approach, to handle the changes in data distribution between the source and the target datasets. While the sample re-weighting technique in TransFall is similar to that of IWLSPC, TransFall performs a two-tier data transformation on both datasets, to empirically match the distributions of the two datasets for more accurate label estimation.

![Figure 4: Results for cross-platform scenario sub-cases.](image)

### Table 1: Dataset and experiment abbreviations.

| Notation | Description |
|----------|-------------|
| Dataset  | Phone Watch HART | 8 smartphones and 9 subjects 4 smartwatches and 5 subjects one smartphone and 30 subjects |
| Cross-Platform | P2P-S P2P-D W2W-S W2W-D | Same-model phone-to-phone Different-model phone-to-phone Same-model watch-to-watch Different-model watch-to-watch |
| Comparison Group | NN DT LR SVM IWLSPC | Nearest neighbor Decision tree Logistic regression Support vector machine Method in [4] |

3.1 Cross-Platform Transfer Learning

This scenario refers to the case when source data and target data are gathered using different devices. We further divide this scenario into four cases as shown in Table 1.

Figure 4 shows the results of label estimation using different approaches in the four sub-cases accordingly, where the red central mark on each box indicates the mean value of labeling accuracy,
the bottom and top edges indicate the 25th and 75th percentiles respectively, and the outliers are denoted as plus symbols.

Overall, all the approaches achieve a higher accuracy on the data gathered by smartphones than that of smartwatches, mainly due to the less stability of data collection using smartwatches. Given the data gathered by smartphones, the labeling accuracy is higher in P2P-S case than that of P2P-D case. This result is coincident with general expectation, because source device and target device are in different models for P2P-D case, which results in large diversity between the obtained datasets due to the differences, such as sampling frequency and platform configuration.

Compared to the other five approaches shown in Figure 4, TransFall achieves the highest labeling accuracy in average, with a correct labeling rate of 0.88, 0.79, 0.41 and 0.56 in the four cases respectively. The increase in the labeling accuracy of TransFall compared to other approaches is > 2.7% for P2P-S, > 6.3% for P2P-D, > 16.6% for W2W-S, and > 16.1% for W2W-D.

In this task, logistic regression model was empirically chosen to carry out activity recognition after label information transfer. Figure 5 shows the results in four transfer learning cases. In general, the classification accuracy is consistent with the labeling accuracy, because the quality of training dataset is determined by the precision of labeling the target data in previous task.

In Figure 5, the performance upper bound (referred to as "Upper") of machine learning model trained with the ground truth appears to be dramatically lower on the data gathered by smartwatches than that of smartphones. The accuracy upper bound is 93.3% in P2P-S case, but 75.9% in W2W-S case. This performance decline reflects the lower quality of sensor data samples collected using smartwatches.

TransFall still achieves improved performance over other approaches. Its classification accuracy is 88.4%, 76.6%, 35.9% and 47.5% for the four sub-cases, respectively. Moreover, the performance improvement of the machine learning model trained by TransFall is > 7.7% compared to the machine learning models trained by other approaches on smartphone data, and > 19% on smartwatch data.

3.2 Cross-Subject Transfer Learning

In this scenario, source data and target data are collected from two subjects using the same type of mobile device. Figure 6 shows the results of label estimation on each dataset separately. Almost all the approaches perform best on the HART dataset than the other two datasets. One possible explanation is less noise in the HART dataset which results in a more informative classification task. Moreover, TransFall performs better than other approaches in the comparison group, with an accuracy increase of > 6.9% on the Phone dataset, > 9.5% on the Watch dataset, and > 3.4% on the HART dataset.

Given label sets obtained using the alternative approaches, we can examine the performance of activity models trained with each label set. Figure 7 shows the activity recognition results. Again, the most approaches perform best on the HART dataset. The empirical upper bound of classification accuracy is 95.9% on the HART dataset, 92.1% on the Phone dataset and 88.0% on the Watch dataset. TransFall still exhibits the best recognition performance, with an accuracy increase of > 7.4% on Phone dataset, > 20.2% on the Watch dataset, and > 4.4% on the HART dataset.

4 CONCLUSION AND FUTURE WORK

TransFall integrates a two-tier data transformation to perform transfer learning. Experimental results demonstrate that TransFall can steadily improve activity recognition accuracy comparing to several alternative approaches. A limitation of TransFall is the assumption of consistent feature alignments between source and target datasets. Our future work involves improving the generalizability of the current framework to handle variations in feature space between source and target datasets.

REFERENCES

[1] Davide Anguita et al. 2013. A Public Domain Dataset for Human Activity Recognition using Smartphones. In ESANN.
[2] Peter Bartlett and Howard Leis. 2008. Representer theorem and kernel examples. (2008).
[3] Hal Daume III. 2004. From Zero to Reproducing Kernel Hilbert Spaces in Twelve Pages or Less. (Mar 2004).
[4] Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. 2012. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. Neurocomputing 80 (2012), 93 – 101. Special Issue on Machine Learning for Signal Processing 2010.

Figure 5: Results for cross-platform scenario sub-cases.

Figure 6: Results of cross-subject activity recognition.

Figure 7: Results of cross-subject activity recognition.
[5] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D.
Lawrence. 2009. Dataset Shift in Machine Learning. The MIT Press.
[6] B. W. Silverman. 1978. Density Ratios, Empirical Likelihood and Cot Death.
Journal of the Royal Statistical Society 27, 1 (1978), 26–33.
[7] Allan Stisen et al. 2015. Smart Devices Are Different: Assessing and Mitigating-
Mobile Sensing Heterogeneities for Activity Recognition. In Proceedings of the
13th ACM Conference on Embedded Networked Sensor Systems. ACM, 127–140.