What to Transfer? High-Level Semantics in Transfer Metric Learning for Action Similarity

Ziad Al-Halah, Lukas Rybok and Rainer Stiefelhagen
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Email: {ziad.al-halah, lukas.rybok, rainer.stiefelhagen}@kit.edu

***Abstract*—Learning from few examples is considered a very challenging task where transfer learning proved to be beneficial. Such a learning framework exploits previous experiences and knowledge to compensate for the lack of training data in a novel domain. Knowledge representation plays a vital role in the type and performance of transfer learning approaches, as well as its robustness against negative transfer effect. This aspect is usually not considered in most of the proposed transfer learning methodologies, where the focus is either on the transfer type or on the representation. In this work, we study the use of various high-level semantics in transfer metric learning. We propose a generic transfer metric learning framework, and analyze the effect of different semantic similarity spaces on transfer type and efficiency against negative transfer. Furthermore, we introduce a hierarchical knowledge representation model based on the embedded structure in the attribute semantic space. The evaluation of the framework on challenging transfer settings in the context of action similarity demonstrates the effectiveness of our approach.**

I. Introduction

Knowledge transfer is the ability to leverage experiences and skills obtained previously via a training process to a new task or domain. This feature is an important characteristic of the learning process of human beings. We do not learn tasks in isolation, rather we try to project the experience we gather through out our lives to facilitate the learning of the new task. The ability to transfer gives us the advantage of an initial high performance and to learn faster when handling a new task while using only few trials (or examples) [1]. Thus, there is a growing interest to mimic this ability in machine learning methods in order to cope with extreme situations where standard learning processes fail or perform poorly. A common scenario where transfer learning proved to be quite beneficial is when training data is scarce or not available (e.g. one- and zero-shot learning [2]–[4]). In such cases, usual machine learning methods can not be applied or they are not able to extract a useful model, and therefore, they will fail to generalize well. Another case is when the data distributions of train and test samples are not similar. This violates the main assumption of many machine learning approaches and results in reduced generalization properties [5].

A knowledge transfer method usually tries to tackle one or more of the following questions [5]: 1) *What to transfer?* This entails the type of knowledge most suitable to be transferred across domains. Hence, an important feature of the transferred knowledge is its ability to encode information that is usable and shareable between tasks. 2) *How to transfer?* The process used to incorporate the transferred knowledge from the source domain in the learning of the target task. 3) *When to transfer?* The source and target tasks might be very different, and transferring knowledge between them may be harmful and hinders the learning of the target task (negative transfer). Thus, it is important to find out when previous experiences are applicable and when they are not.

In this work, we focus on the type of information to transfer across domains, hence answering the question: *What to transfer?* and its consequent effect on other transfer options. There are three common approaches in this direction. a) *Feature representation transfer*, where the focus is on learning a good knowledge representation model for the target domain based on relevant information in the source domain [6], [7]. b) *Parameter transfer*, here, the models (or parameters) learned in the source domain are used to regularize or to include as a prior in the model learning of the target task [8]. c) *Instance transfer*, where all or some of the samples in the source domain are re-used in the learning of the target task in order to overcome the low number of target training samples [9]. While most of the previous works tackle these options separately (e.g. [4], [9], [10]), we believe they should be considered jointly. The choice of feature representation and how the knowledge is modeled will influence the efficiency of all transfer options: the representation, instance and parameter transfer. For example, learning the color distributions in a set of animal classes is considered a low-level meta-information that will not hold true when considering different animal categories. However, learning the visual semantic attribute distribution or the meta-relations between the categories would hold true even when moving to a different domain. Such high-level semantics are less influenced by the low-level feature distribution and consequently are an adequate knowledge to be transferred across domains. In transfer metric learning literature, this observation is usually ignored and the focus is on parameter transfer while using low-level knowledge representation [10], [11].

Furthermore, the common assumption in evaluating transfer learning methods is that the source data set is much larger and more diverse than the target set [3], [4], [12]. However, collecting data and labeling are expensive tasks. This will usually result in small datasets for training, and the number of defined categories in source is much less than the expected "unseen" categories in target (i.e. small and simple source domain versus large and diverse target domain). Such an evaluation setup, that we address in this work, imposes a great challenge to transfer learning approaches since they have a limited knowledge source to generalize from it to the target.

To this end, our contribution in this work is: 1) we show the benefits of using high-level semantic representation for transfer metric learning. 2) We propose a novel hierarchical knowledge representation that encodes the embedded semantic structure

of category similarities in the attribute space, and show its superior performance to other semantic models. Furthermore, 3) we introduce a generic framework for representation transfer that improves the metric learning model and reduces the negative transfer effect. 4) The evaluation is conducted in challenging and realistic settings, where the target set is much more diverse and different than the source set.

II. Related Work

Knowledge transfer has attracted a lot of attention in the last years, and several approaches were proposed in various fields. We refer in this section to two closely related sub-fields to our work: the knowledge representation transfer and transfer metric learning. Recent comprehensive surveys on transfer learning can be found in [5], [13].

Transfer metric learning. While standard supervised and semi-supervised metric learning are widely popular [14], only few works tackle the problem of knowledge transfer in metric learning. In [11], the authors integrated multiple source metrics into a regularized metric learning framework, and similar to [15], they used the log-determinant regularization to minimize the divergence between the source metrics and the target metric. In contrast, Zhang and Yeung in [10] considered the transfer metric learning (TML) as a special case of multi-task learning. They jointly learn the relations between the source tasks and the target task while learning the target metric matrix. Their approach, unlike [11], can model positive, negative and zero task correlations. TML showed superior performance to [11] when the training data is scarce [10]. Nonetheless, both approaches used parameter transfer based on low-level feature representation. To the best of our knowledge, the use of high-level semantics and the effect of knowledge representation on other types of transfer options were not addressed before in the context of transfer metric learning.

Knowledge representation transfer. Various knowledge models were introduced in the literature. A common representation that gained a lot of attention recently is semantic attributes. They describe the visual appearance of an entity and represent an intermediate semantic layer between the low-level features and categories. Attributes were successfully used in transfer learning applications like object and action zero-shot recognition [3], [4], [16]. Another model was introduced by Bart and Ullman [2]. They represented an instance of an unseen class by its similarity to known categories for one-shot object recognition, and showed a significant improvement in classification performance. Hierarchies, on the other hand, are considered a popular knowledge representation. In contrast to other types of representations, they are able to capture information at different resolution levels. Usually, the structure is defined either manually [17], based on external linguistic sources like WordNet [12] or automatically driven from data [18]. Still, linguistic sources are not suitable for all types of information. For example, visual relations between actions are not well defined in such sources since actions are mapped to verbs and not nouns. On the other hand, constructing a hierarchy based on low-level data distribution is not favorable in transfer settings, since source and target may differ significantly in this regard. To this end, and to avoid using manually tuned hierarchies, we introduce a hierarchical model that is learned from the embedded structure in the attribute space and encodes the relative similarities between categories.

Additionally, in the evaluation of transfer learning approaches, it is commonly assumed that the target set is smaller and less diverse than the source set [2]–[4], [10], [16]. Even in the large-scale evaluation of [12], where different transfer approaches for zero-shot object recognition are tested, the source set contained four times more classes than the target set. We address here the opposite settings, i.e. the source of knowledge is smaller and less diverse than the target set, which represent a more challenging evaluation settings.

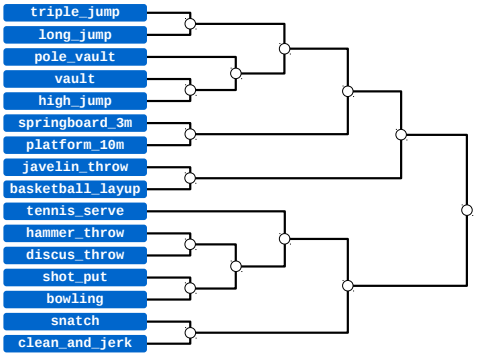


Fig. 1: The learned hierarchy of action classes in Olympic Sports.

III. Approach

A. Semantic Similarity Space

Most of previous approaches use low-level features to compare objects or actions [10], [11], [15], [19]. We believe that semantics at different levels of complexity can be a better representation of source knowledge to be transferred across domains. While the feature similarity space is usually high dimensional and dependent on the data distribution in the source domain, the semantic similarity space is lower dimensional, concise and is more robust to the changes in the data distributions between target and source domains. There are two common semantic spaces that are usually used as an intermediate representation. The *attribute similarity space*, where instances are represented by their visual properties, and the *category similarity space*, where instances are represented by their resemblance to other previously learned categories. We also introduce a novel representation of a third similarity space, the *hierarchical similarity space*, where instances are represented by a hierarchical structure that captures the visual properties of the instance at different resolution levels.

1) ***Attribute Similarity Space****:* Attributes define an intermediate representation between low-level features and high level categories [3], [4]. Semantic attributes describe an entity regarding its visual appearance, parts and motion patterns (e.g. *is-round*, *has-ears* and *forward-motion*). Hence, they can be easily shared across categories and even used to predict unseen classes if they can be described using the same set of attributes.

In the attribute similarity space $\mathcal{A}$, the different semantic attributes span the basis of the space where each axis encodes the presence of one of the attributes as well as its intensity (or confidence for binary attributes) in a certain data instance. Samples that belong to the same category are close to each other in $\mathcal{A}$ since they share the same properties, and they will form a tight cluster of points that are distinguishable from other samples of different categories. Therefore, the closer the points are to each other in $\mathcal{A}$ the more attributes they share, and consequently, the more similar they are.

The samples in the d dimensional feature space $\mathcal{X}^d$ are

mapped to space $\mathcal{A}$ using $f_{\mathcal{A}}(x)$:

$$\begin{array}{rcl} f_{\mathcal{A}}(x) & : & \mathcal{X}^d \rightarrow \mathcal{A}^n \quad \text{and} \\ f_{\mathcal{A}}(x) & = & [f_{a_1}(x), f_{a_2}(x), \ldots, f_{a_n}(x)]^T \end{array} \tag{1}$$

where $f_{a_i}(x)$ is the prediction score of attribute a_i on instance x, and n is the number of defined attributes.

*2) **Category Similarity Space**:* Humans do not only use visual properties to describe entities in their environment. It is also common to use inter-class relations as means of description. Consider for example the action class *triple-jump*; it can be described as an action similar to class *run* and class *jump*. This intra-class similarity pattern is not specific to a certain sample of *triple-jump*, rather it characterizes all samples that belong to this category.

In that sense, the category similarity space $\mathcal{C}$ provides a meaningful semantic space to compare different actions in terms of their similarity patterns to previously learned categories [2]. In $\mathcal{C}$, the bases are spanned by the predefined categories, where each axis encode the resemblance of a sample to a learned category.

Samples from the feature space are mapped to $\mathcal{C}$ based on $f_{\mathcal{C}}(x)$:

$$\begin{array}{rcl} f_{\mathcal{C}}(x) & : & \mathcal{X}^d \rightarrow \mathcal{C}^m \quad \text{and} \\ f_{\mathcal{C}}(x) & = & [f_{c_1}(x), f_{c_2}(x), \ldots, f_{c_m}(x)]^T \end{array} \tag{2}$$

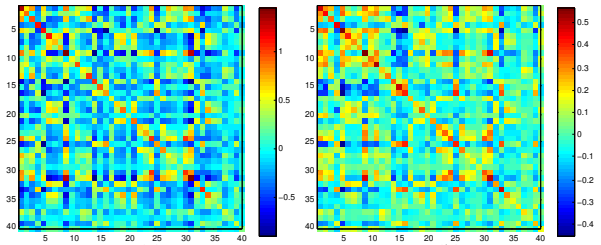
where $f_{c_i}(x)$ is the prediction score of category c_i on instance x and m is the number of categories.

*3) **Hierarchical Similarity Space**:* A common property of the previously defined spaces is that both of them represent semantics at a single layer of resolution. That is, both of them ignore the implicit structure that exists in the semantic space. Such structure allows us to have semantics depicted at various levels of resolution or complexity, which enriches the knowledge obtained in the source domain and provides a better semantic representation of samples.

Consider for example the categories *walk*, *jump* and *jump-forward*. While the last one is partially similar to the first two, it is better represented by the combination of both, *walk-jump*, and learning the common pattern between these two classes can provide a higher category of semantics that suits the novel class of *jump-forward*.

However, constructing such a hierarchy based on low-level features will not necessarily result in a semantically meaningful structure. Hence, we propose to capture this hierarchical model by exploiting the similarity between categories in the attribute space. Attributes correspond to observable properties of the categories, and the more attributes are shared between a couple of categories the higher is the overall visual similarity between the pairs.

Based on this observation, and assuming that each of the action classes is described with a vector of semantic attributes of length n, we can exploit this representation in order to build our hierarchical model by grouping action categories close to each other in the attribute space (Figure 1). This is achieved by applying an agglomerative hierarchical clustering algorithm over the attributes representation of the classes to get a dendrogram depicting the hierarchical clustering result. The dendrogram is then used to construct the final actions hierarchy by interpreting the action classes as leaf nodes and the intermediate clusters at different similarity threshold levels as inner nodes. The sub- and super-cluster relations are translated to *is-a* relations in the tree structure. For the case of



(a) (b)

Fig. 2: The attribute correlations on a) source set (Olympic Sports) and b) target set (ASLAN), (Best seen in color).

using binary attributes to describe the various action classes, we use a hierarchical k-means clustering algorithm with the Manhattan distance (L_1) to capture the similarity in $\mathcal{A}^n$.

Then instances $x \in \mathcal{X}^d$ are represented in the hierarchical similarity space $\mathcal{H}$:

$$\begin{array}{rcl} f_{\mathcal{H}}(x) & : & \mathcal{X}^d \rightarrow \mathcal{H}^k \quad \text{and} \\ f_{\mathcal{H}}(x) & = & [f_{n_1}(x), f_{n_2}(x), \ldots, f_{n_k}(x)]^T \end{array} \tag{3}$$

where $f_{n_i}(x)$ is the prediction score of node i in the hierarchy, and k is the number of nodes. The node classifiers are trained in child-vs-parent manner. That is, if $\text{pos}(n_i) = \bigcup \text{pos}(n_j)$ is the positive set of node n_i where $n_j \in \text{child}(n_i)$, then the classifier f_{n_i} is trained on $\text{pos}(n_i)$ as the positive set against $\{\text{pos}(n_p)/\text{pos}(n_i)\}$ as the negative set, where $n_p = \text{parent}(n_i)$.

B. Decorrelated Normalized Space

The learned semantic similarity spaces will implicitly model the correlations of the data in the training set. Such correlations are related to the data distribution in the source domain which most likely differ significantly from the distribution in the target domain (Figure 2). Hence, transferring such knowledge across domains will likely result in a negative transfer effect [1], [5]. Therefore, it is quite important to eliminate the correlations learned in the source domain from the semantic spaces in order to restrict the negative transfer.

Motivated by the work of [20] on attribute decorrelation and [21] on removing co-occurrence patterns from the bag-of-words model, the decorrelation of the semantic similarity space $\mathcal{S}$ ($\mathcal{S} \in \{\mathcal{A}, \mathcal{C}, \mathcal{H}\}$) can be efficiently achieved using the whitening transformation. Considering the data in $\mathcal{S}$ is represented by matrix $\mathbf{Y}$, then the correlations are modeled by the covariance matrix $\mathbf{\Omega} = \mathbf{Y}\mathbf{Y}^T$. By whitening $\mathbf{Y}$, the data is transformed to space $\tilde{\mathcal{S}}$ where the bases are decorrelated and given same importance, which is the result of transforming $\mathbf{\Omega}$ to the identity matrix. The whitening transformation $\mathbf{W}$ of $\mathcal{S}$ is obtained by analyzing the covariance matrix $\mathbf{\Omega}$ such that:

$$\mathbf{W} = \mathbf{V}\mathbf{\Sigma}^{-1/2} \quad \text{and} \quad \mathbf{\Omega} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T \tag{4}$$

The columns of $\mathbf{V}$ are the eigenvectors of the covariance matrix, and $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues ($\mathbf{\Sigma}_{ii} = \lambda_i$). If some eigenvalues are very small ($\lambda_i < \theta$) we ignore the corresponding vectors in $\mathbf{V}$ to have a robust estimation of $\mathbf{W}$:

$$\hat{\mathbf{W}} = \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{-1/2} \quad \text{where} \quad \hat{\mathbf{\Sigma}}_{ii} > \theta \tag{5}$$

The vectors in the whitened space are then normalized by their norm to have a better estimation of the similarity. Thus,

the data representation in $\mathcal{S}$ is transformed to the decorrelated normalized space $\mathcal{S}_{dn}$ using:

$$f_{\mathcal{S}_{dn}}(x) = \hat{\mathbf{W}}^T y / \|\hat{\mathbf{W}}^T y\|_2 \quad \text{where} \quad y = f_{\mathcal{S}}(x) \tag{6}$$

C. Similarity Metric Learning

To compare instances in the semantic similarity space, we learn a similarity metric in that space in order to adapt to the positive and negative pairs distribution in the target domain. For that purpose we use the Logistic Discriminant based Metric Learning (LDML) [19].

The metric learning problem in LDML is formulated as a standard logistic discriminant model where the maximum log-likelihood is used to optimize the parameters of the model. LDML has a convex optimization objective which guarantees an optimum global solution. However, our approach is not restricted to a certain metric learning method as we will show later in Section IV-D.

IV. Evaluation

We evaluate our framework using two publicly available data sets. The first one is the Olympic Sports data set [22]. It contains 781 videos of 16 action classes collected from YouTube, like *hammer-throw*, *tennis-serve* and *triple-jump*. We use the attribute annotations provided by [16], where the actions are labeled with 40 semantic attributes describing motion, pose and objects, such as *lift-something*, *throw-away*, *two-arms-open* and *outdoor*. The second data set is ASLAN which has been recently published in [23]. The data set is collected for the main task of comparing actions (similar/not-similar). It has about 432 action classes with more than 3600 video samples. Each class has about 8.5 video samples with more than 100 classes having only one sample each.

In our experimental settings, we use Olympic Sports as source and ASLAN as target data set. This poses a very challenging problem because of the high diversity in ASLAN compared to Olympic Sports (432 to 16 different classes).

As a video descriptor, we use the bag-of-words (BoW) model based on histograms of oriented gradients and optical flow (HOGHOF) [24] with a vocabulary of size 4000. We use that BoW model to train the different classifiers, presented in Section III-A, on the training split of Olympic Sports. The features are preprocessed with a power transform [25] ($\alpha = 0.3$) before training a linear support vector machine, where the parameters of the classifiers are estimated using a 5-fold cross validation. For the decorrelated normalized space, we set $\theta = 10^{-8}$. To simulate a real transfer learning problem, we do no further training of classifiers or the BoW model on the target set (ASLAN), and only the similarity metric is adapted from the available training data to infer a reasonable comparison metric in each of the semantic similarity spaces. The threshold of similarity is automatically learned using a linear SVM trained on the distances between training pairs.

A. Knowledge Representation Transfer

We test first the performance of different semantic spaces compared to the common low-level similarity space. We learn the different knowledge representations on Olympic Sports then we evaluate on ASLAN using the view 1 training/testing split [23]. The number of training pairs of similar and dissimilar actions is varied from 5% to 100% of the training set. For each run, a random subset of the training pairs is selected to learn the similarity and then evaluated on the test split. This is repeated 10 times, and we report the average accuracy and

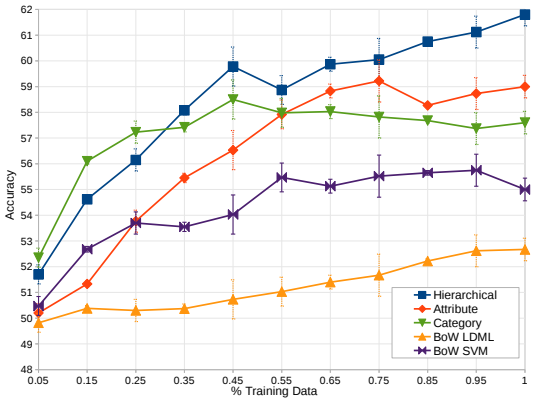


Fig. 3: Overall performance of different semantic similarity spaces regarding various sizes of the target training set.

standard error of similarity classification as seen in Figure 3. For the feature space, we report two methods: the first is using LDML after reducing the dimensionality of the features to 128 using principle component analysis (it is intractable to use the full feature vector with LDML [19]). While in the second we use the full feature vector (4000) and train an SVM on the element wise multiplication of the training pairs $[x_1 . * x_2]$ (using the absolute difference $|x_1 - x_2|$ or the concatenation of the previous two produced inferior performance).

From Figure 3, we see that the three semantic spaces outperform the low-level feature space. The hierarchical and category similarity spaces outperform the attribute space when the training data is scarce. However, when more than half of the training data is available, the attribute space seems to do better than the category space while the proposed hierarchical model outperform both. This confirms our previous hypothesis on the importance of high-level semantics and their ability to generalize well when transferred to other domains.

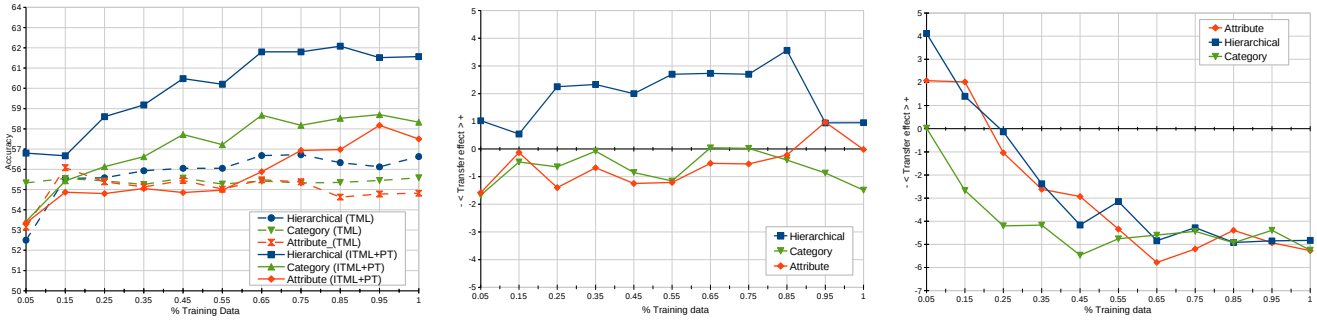
B. Parameter Transfer

The similarity metric learning method LDML does not allow for parameter transfer in its formulation. Hence we propose instead a simple parameter transfer approach based on the information-theoretic metric learning (ITML) [15]. The metric learning problem in ITML is defined as:

$$\min_{\mathbf{M}} \quad \mathrm{KL}(p(x, \mathbf{M}_0) \parallel p(x, \mathbf{M})), \tag{7}$$

where KL is the Kullback-Leibler divergence between two Gaussian distributions corresponding to a prior metric $\mathbf{M}_0$ and the learned metric $\mathbf{M}$. Additionally, some few constraints on the upper and lower bound of distances between similar and dissimilar pairs are taken into account [15]. In (7), the prior $\mathbf{M}_0$ is usually set to the identity matrix $\mathbf{I}$ (Euclidean metric) or the inverse of the covariance matrix. In contrast, we suggest a parameter transfer approach by setting the prior to be the metric learned in the source data set ($\mathbf{M}_0 = \mathbf{M}_{source}$). In this case, the metric learning in the target set is regularized to be close to the source metric ($\mathbf{M}_{source}$) while at the same time satisfying the constraints on the pair distances in the target set.

We evaluate the parameter transfer setting by learning first the similarity metric for each of the three semantic spaces (Section III-A) in the source set (Olympic Sports) and transfer that metric using (7) to the target set (ASLAN). In order to learn the transferred metric matrix, we randomly generate 1500



(a) (b) (c)

Fig. 4: (a) Comparison of the proposed parameter transfer approach ITML+PT to TML and the robustness of different knowledge representations to negative transfer effect using (b) parameter and (c) instance transfer.

pairs of similar and dissimilar actions in the source, and learn the similarity using the standard proposed framework. During testing, we use the same settings as described in Section IV-A.

We compare ITML with the proposed parameter transfer approach (ITML+PT) to state-of-the-art transfer metric learning (TML) [10]. We set the parameters for both ITML and TML as suggested by the authors in [15] and [10], respectively. Interestingly, ITML+PT outperforms TML (Figure 4a). TML seems to have a saturated performance after using just 15% of the training set and slightly profits from the different semantic representations, while ITML+PT has a higher initial performance and clearly takes advantage of the characteristics of the different similarity spaces. This can be due to the formulation of TML as a special case of multi-task metric learning, and the assumption that the tasks (source and target) share a common data distribution which is not the case here.

We also analyze the transfer effect (positive or negative) as the difference in performance (measured by accuracy) between using the parameter transfer and without (i.e. setting $\mathbf{M}_0 = \mathbf{I}$ in (7)). Both attribute and category similarity spaces show a negative transfer effect while the hierarchical space benefits from parameter transfer (Figure 4b). It seems, as motivated in Section I, that the higher the level of semantic knowledge encoded in the model, the more robust is the model against negative transfer. Since the learned meta-information (parameters) in source domain can still be true in the target even though they have very different data distributions.

C. Instance Transfer

Here, we add the training pairs from the Olympic Sports to the training set in ASLAN and evaluate using varying sizes of the target train set. We report the difference in performance as in Section IV-B.

We see in Figure 4c that while the hierarchical and attribute spaces take advantage of the additional samples when the training set is extremely small (less than 25%), the instance transfer produces a negative transfer for all semantic spaces otherwise. This type of transfer introduces an extreme change in the data distribution of the target train set which is not reflected in the test set, resulting in performance deterioration. This also shows how the target and source sets are different and how challenging are the transfer settings.

D. Importance of Decorrelated Normalized Space

In this experiment, we test the contribution of the proposed decorrelated-normalized space (DNS) to the transfer performance. We test using three state-of-the-art metric learning methods (ITML [15], LDML [19] and KISSME [26]) and three commonly used metrics (the Mahalanobis distance using the inverse of the covariance (Cov^{-1}), the L_2, and SVM as used in Section IV-A). We use all training pairs in the target and report the accuracy with and without using a decorrelated normalized space.

TABLE I: The effect of the decorrelated normalized space (DNS) on the performance of various state-of-the-art metric learning methods.

Space / Metric	ITML [15]	LDML [19]	KISSME [26]	Cov^{-1}	L_2	SVM
$\mathcal{H}$	58.38	54.23	55.50	51.03	52.85	**57.58**
$\mathcal{H}_{dns}$	**60.62**	**61.80**	**60.98**	**56.98**	**56.33**	56.90
$\mathcal{A}$	55.08	57.80	55.50	50.87	54.00	57.50
$\mathcal{A}_{dns}$	**57.52**	**59.00**	**58.42**	**56.37**	**54.83**	**57.73**
$\mathcal{C}$	57.65	57.50	56.77	54.17	54.50	53.17
$\mathcal{C}_{dns}$	**59.82**	**57.60**	**61.20**	**57.23**	**55.50**	**57.63**
$\mathcal{X}$	55.38	**58.95**	49.83	49.67	50.00	50.00
$\mathcal{X}_{dns}$	**56.07**	52.67	**54.33**	**53.00**	**56.53**	**56.05**

In Table I, we see that in most of the cases (22 out of 24), the decorrelated space increased the performance of the transfer metric (up to 7% absolute increase). We also see that the proposed framework is quite generic and not restricted to a certain metric learning method. Even when using a simple metric as L_2, DNS helped to learn a better similarity metric. The hierarchical model achieves the best performance using LDML with 61.80% and both the hierarchical and category space seems to do better than the attribute model.

E. Full Scale Evaluation

We compare the performance of the proposed transfer metric learning framework with standard metric learning methods when the train data is abundant (i.e. the knowledge representation is learned in target set and no transfer learning is carried out). This is a widely ignored evaluation setting in transfer learning publications where the focus is only on the case when the training data is scarce. Evaluating on the large scale data set helps us to put the transfer metric learning method in perspective to other methods that have the advantage to adapt well to the target data distribution.

For that purpose, we use ASLAN view 2 which has 6000 pairs of similar and dissimilar actions, and we report the performance in terms of accuracy and area under receiver operating characteristic (ROC) curve and using 10-fold cross

TABLE II: Large scale evaluation on view 2 of the ASLAN data set.

Representation Learning in Source	$\mathcal{H}_{dns}$	$\mathcal{C}_{dns}$	$\mathcal{A}_{dns}$	$\mathcal{X}_{dns}$
#Dimension	30	16	40	128
LDML	$\mathbf{59.18 \pm 0.98 (62.16)}$	$57.85 \pm 1.02 (60.57)$	$57.30 \pm 0.58 (60.85)$	$56.97 \pm 0.69 (60.15)$
Representation Learning in Target	**HOG**	**HOF**	**HNF**	**HOG+HOF+HNF**
#Dimension	5000	5000	5000	3 x 5000
$\sqrt{\sum(x_1 . * x_2)}$	$58.55 \pm 0.80 (61.59)$	$56.82 \pm 0.57 (58.56)$	$\mathbf{58.87 \pm 0.89 (62.16)}$	$60.08 \pm 1.08 (63.89)$
Hellinger	$53.22 \pm 0.61 (54.19)$	$53.77 \pm 0.72 (56.00)$	$53.77 \pm 0.73 (55.80)$	$54.83 \pm 0.90 (57.18)$
Chi-Square	$53.28 \pm 0.69 (54.42)$	$53.42 \pm 0.62 (55.79)$	$53.87 \pm 0.72 (55.97)$	$54.97 \pm 0.97 (57.13)$
12 Similarities	$59.78 \pm 0.82 (63.20)$	$56.68 \pm 0.56 (58.97)$	$59.47 \pm 0.66 (63.30)$	$\mathbf{60.88 \pm 0.77 (65.30)}$

validation as suggested in [23]. For an in-target representation modeling, we compare to Kliper-Gross *et al.* [23] approach. They propose to learn a BoW model of size 5000 for each of the three features HOG, HOF, and HNF [24] to represent the actions. They use 12 different similarity metrics to compare actions based on each of these representations and a combination of the three. We report in Table II the results of their best single similarity metric and the results of using the combination of the 12 metrics as stated in [23].

We see in Table II that the performance of the different semantic spaces in the transfer metric approach follows the complexity level of semantics encoded in the model, with the proposed hierarchical representation doing best, followed by the category and attribute spaces. More importantly, the transfer metric method performs as well on the target set as the approach that uses a representation learned in target domain. Even when 12 different similarities and 3 feature representations are combined; the gain in performance of the in-target method is 1.7% in accuracy. This is an impressive performance for the transfer metric learning approach, bearing in mind the diversity of the target compared to the source set (432 to 16 classes) and that the data representation learned in the source was never adapted to model changes in the target domain.

V. Conclusion

We proposed a generic framework for transfer metric learning and showed the importance of knowledge representation on different transfer options. High-level semantics have better transfer properties and encode richer transferable knowledge in comparison to low-level features. We introduced a hierarchical representation that models the embedded structure of category similarities in the attribute space. The proposed hierarchical model performed best and was more resilient to negative transfer effect. In addition, different metric learning methods benefit from the proposed transfer framework. We evaluated on very challenging settings where the target set is much more complex and diverse in comparison to the source set. Nonetheless, we showed that even when the knowledge source is limited, transfer learning can still be beneficial if an appropriate semantic representation is used. Finally, large-scale evaluation showed impressive results of the transfer approach; the performance is in line with methods using in target feature representation learning.

Acknowledgment

This study is funded by OSEO, French State agency for innovation, as part of the Quaero Programme.

References

[1] L. Torrey and J. Shavlik, "Transfer Learning," *Handbook of Research on Machine Learning*, 2009.

[2] E. Bart and S. Ullman, "Single-example learning of novel classes using representation by similarity," in *BMVC*, 2005.

[3] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing Objects by their Attributes," in *CVPR*, 2009.

[4] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009.

[5] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *TKDE*, vol. 22, pp. 1345–1359, 2010.

[6] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-View Action Recognition via View Knowledge Transfer," in *CVPR*, 2011.

[7] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis." *NNLS*, vol. 22, pp. 199–210, 2011.

[8] F. Nater, T. Tommasi, H. Grabner, L. V. Gool, and B. Caputo, "Transferring Activities : Updating Human Behavior Analysis," in *ICCV Workshop on Visual Surveillance*, 2011.

[9] A. Lam, A. K. Roy-Chowdhury, and C. R. Shelton, "Interactive Event Search Through Transfer Learning," in *ACCV*, 2010.

[10] Y. Zhang and D.-Y. Yeung, "Transfer metric learning by learning task relationships," in *ACM SIGKDD - KDD*, 2010.

[11] Z.-J. Zha, T. Mei, M. Wang, Z. Wang, and X.-S. Hua, "Robust Distance Metric Learning with Auxiliary Knowledge," in *International Joint Conference on Artificial Intelligence*, 2009.

[12] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting," in *CVPR*, 2011.

[13] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer Learning for Activity Recognition: A Survey." *KAIS*, vol. 36, pp. 537–556, 2013.

[14] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," Tech. Rep., 2013.

[15] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[16] J. Liu, B. Kuipers, and S. Savarese, "Recognizing Human Actions by Attributes," in *CVPR*, 2011.

[17] A. Zweig and D. Weinshall, "Exploiting Object Hierarchy: Combining Models from Different Category Levels," in *ICCV*, 2007.

[18] R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "Learning to Share Visual Appearance for Multiclass Object Detection," in *CVPR*, 2010.

[19] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric Learning Approaches for Face Identification," in *ICCV*, 2009.

[20] Z. Al-Halah, T. Gehrig, and R. Stiefelhagen, "Learning Semantic Attributes via a Common Latent Space," in *VISAPP*, 2014.

[21] H. Jegou and O. Chum, "Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening," in *ECCV*, 2012.

[22] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," in *ECCV*, 2010.

[23] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge." *T-PAMI*, 2012.

[24] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[25] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012.

[26] K. Martin, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large Scale Metric Learning from Equivalence Constraints," in *CVPR*, 2012.