Pattern Recognition Letters
journal homepage: www.elsevier.com

Transfer Metric Learning for Action Similarity using High-Level Semantics

Ziad Al-Halah[a,**], Lukas Rybok[a], Rainer Stiefelhagen[a]

[a] *Karlsruhe Institute of Technology, Institute for Anthropomatics and Robotics, Karlsruhe 76131, Germany*

ABSTRACT

The goal of transfer learning is to exploit previous experiences and knowledge in order to improve learning in a novel domain. This is especially beneficial for the challenging task of learning classifiers that generalize well when only few training examples are available. In such a case, knowledge transfer methods can help to compensate for the lack of data. The performance and robustness against negative transfer of these approaches is influenced by the interdependence between knowledge representation and transfer type. However, this important point is usually neglected in the literature; instead the focus lies on either of the two aspects. In contrast, we study in this work the effect of various high-level semantic knowledge representations on different transfer types in a novel generic transfer metric learning framework. Furthermore, we introduce a hierarchical knowledge representation model based on the embedded structure in the semantic attribute space. The evaluation of the framework on challenging transfer settings in the context of action similarity demonstrates the effectiveness of our approach compared to state-of-the-art.

1. Introduction

Instead of learning new concepts in isolation, humans have the ability to consider connections to previously obtained skills and experiences, which makes our learning process extremely efficient (Reder and Klatzky (1994)). In psychology, this skill is known as *knowledge transfer* or *transfer learning* (Woodworth and Thorndike (1901)). It gives us humans the advantage of learning new concepts faster and with a high initial performance when using only few trials or examples (Torrey and Shavlik (2009)). In contrast, most machine learning algorithms require a large number of training examples, since training only relies on domain specific data, instead of incorporating prior knowledge (Fei-Fei (2006)). However, in cases when training data is scarce or not available, such methods can not be applied or are unable to extract a useful model, and thus fail to generalize well. Therefore, there is a growing interest in the Machine Learning Community to mimic this human ability. A typical task that benefits from knowledge transfer is one- and zero-shot learning (Bart and Ullman (2005); Farhadi et al. (2009); Lampert et al. (2009)).

Another problem of many machine learning models is their assumption that training samples are drawn according to the same probability distribution as the unseen test samples (Valiant (1984)). Nevertheless, this hypothesis does not always hold in practical problems, resulting in a reduction of generalization properties. For instance, consider you have built a robust classifier to distinguish between different sports actions and would like to use the same system on more general videos found on YouTube. Usually, this would require an expensive data collection and annotation process. However, using transfer learning methods, it is possible to re-use an established model to save a significant amount of labeling effort (Pan and Yang (2010)).

According to Pan and Yang (2010), transfer learning research tries to solve one or more of the following three problems:

1. *"What to transfer?"*, asks what type of knowledge representation is 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?"*, asks how the transferred knowledge from the source domain can be incorporated in the learning of the target task.
3. *"When to transfer?"*, asks when transfer learning is beneficial, since knowledge transfer can sometimes decrease

[**] Corresponding author: Tel.: +4972160844735; Fax: +4972160845939; *e-mail:* ziad.al-halah@kit.edu (Ziad Al-Halah)

the effectiveness of learning in the target domain (negative transfer). This can for instance happen, when the source and target tasks are very different.

The focus of our work lies on the type of information to transfer across domains, hence on answering the question: *What to transfer?* and consequently of its effect on different types of transfer methods. There are three common approaches in that direction:

1. *Feature representation transfer*, where a knowledge representation model is learned or adopted for the target domain, based on relevant information in the source domain (Liu et al. (2011b); Pan et al. (2011)).
2. *Parameter transfer*, where models (or parameters) are learned in the source domain and then used to regularize or to be included as a prior in the model learning of the target task (Nater et al. (2011)).
3. *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 (Lam et al. (2010)).

Unlike previous works, which analyze these transfer types separately (e.g. Lampert et al. (2009); Zhang and Yeung (2010); Lam et al. (2010)), we believe that they should be considered jointly. The choice of the feature representation and how the knowledge is modeled will eventually influence the efficiency of all three approaches: the representation-, instance- and parameter transfer. For example, when using color distributions learned on sea animals as a low-level representation, this most likely will generalize poorly to a domain of bird categories and result in a bad performance. However, learning the meta-relations between the categories or the visual semantic attributes (e.g. *has-head*, *is-round* and *has-stripes*) would result in constructing a knowledge space that can be easily shared between various domains. Such high-level semantics are less likely to be influenced by the low-level feature distribution, and consequently form an adequate knowledge representation to be transferred across domains. In transfer metric learning literature, this observation is usually ignored and instead the focus lies on parameter transfer while only using a low-level knowledge representation (Zhang and Yeung (2010); Zha et al. (2009)).

Another common assumption in the transfer learning literature is that the source data set is much more diverse and complex than the target set and thus the experimental evaluation protocol is designed accordingly (e.g. Farhadi et al. (2009); Lampert et al. (2009); Rohrbach et al. (2011)). However, collecting and annotating new data is an expensive effort. While we might create data sets of hundreds of action categories there is still tens of thousands of "unseen" classes (i.e. with no training examples). Hence, it seems that it is more likely that we will have a small and simple source domain against a large and diverse target domain. Moreover, the usual case in most research fields is to first focus on solving simple problems before moving on to more complex ones. For instance, the action recognition community started with the task of classifying simple actions in controlled environments (e.g. Schüldt et al. (2004)) and then slowly moved to the complex Action Similarity Labeling (ASLAN) Challenge proposed by Kliper-Gross et al. (2012), and beyond. Thus, it would be beneficial if each time we switch to a more challenging task, all previously collected data and experience could be successfully used to improve task performance in the new complex domain. Therefore, we address in our work an evaluation setup where the number and complexity of categories in the source domain is much lower than in the target domain. Such a setup imposes a greater challenge to transfer learning approaches.

In conclusion, the contribution of our work is as follows:

- We show the benefits of using high-level semantics for transfer metric learning.
- 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.
- We introduce a novel generic framework for transfer metric learning that improves the transfer performance and reduces the negative transfer effect.
- We suggest a realistic and challenging evaluation protocol for transfer learning, where the target domain is much more diverse and complex than the source domain.

This work is an extended version of Al-Halah et al. (2014b). The main additional contribution is an extended evaluation, and discussion of the results. In the added experiments, the emphasis lies on the analysis of how the knowledge complexity of the source sets affects the transfer process.

2. Related Work

Transfer learning has attracted a lot of attention in the last years, and several approaches were proposed in various fields. Since, it is out of scope of this work to summarize all past research efforts, we refer the interested readers to the comprehensive surveys by Pan and Yang (2010), and Cook et al. (2013), and focus on the most related sub-fields.

Transfer metric learning
While standard supervised and semi-supervised metric learning are widely popular (cf. the survey by Bellet et al. (2013)), to the best of our knowledge only two works exist that analyze the application of metric learning to the problem of knowledge transfer. Zha et al. (2009) propose to integrate multiple source metrics into a regularized metric learning framework and make use of log-determinant regularization to minimize the divergence between the source metrics and the target metric. A drawback of this approach is that it can only represent positive and zero task correlation, but not negative task correlations. Therefore, Zhang and Yeung (2010) proposed a unified framework, called Transfer Metric Learning (TML), that models all three task correlations, while also guaranteeing to find a globally optimal solution. TML is formulated as a special case of multi-task learning, where several independent source tasks and one target task are given, and the relations between the sources

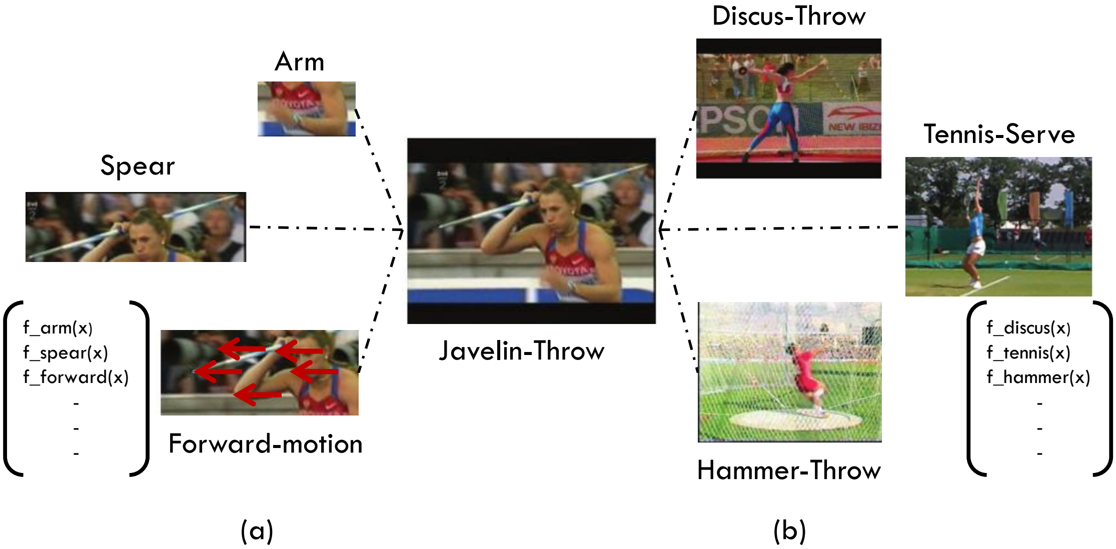


Fig. 1: The attribute-based representation (a) captures some fine-grained visual properties of an action, such as motion pattern, body parts, and objects; while the category-based representation (b) encodes the overall similarity of a certain action to the various categories.

and the target are jointly modeled when learning the target metric matrix. Compared to the work of Zha et al. (2009), TML showed a superior performance when the training data is scarce. Nonetheless, unlike our work, both approaches use parameter transfer based solely on a low-level feature representation. To the best of our knowledge, the use of high-level semantics and the analysis of the impact of different knowledge representations on the different transfer types have not been addressed before in the context of transfer metric learning.

Knowledge representation transfer
Most of the previous work tackles the distribution differences between the source and target domain as a domain adaptation problem of the low-level features (Pan et al. (2011); Gong et al. (2012)) or by learning a robust and transferable sparse representation (Long et al. (2013)). In contrast to this line of research, we study in this work the robustness of high-level semantic representations in challenging transfer settings. Unlike the common case of domain adaption, transferring high-level knowledge representation does not require the availability of target data at time of representation learning which facilitates and generalizes the transfer process. Moreover, as we will show later in the evaluation, high-level semantics exhibit better performance when transferred across data sets compared to low-level features.

Among the various knowledge models that were introduced recently in the literature, semantic attributes have gained an increasing amount of attention. They describe the visual appearance of an entity and represent an intermediate semantic layer between the low-level features and class categories. Attributes were successfully used in transfer learning applications, like zero-shot recognition of objects, and actions (Lampert et al. (2009); Farhadi et al. (2009); Liu et al. (2011a)). Another approach to represent an instance of an unseen class is by its similarity to known categories. This has been applied by Bart and Ullman (2005) to one-shot object recognition resulting in a significant improvement in classification performance compared to low-level features.

On the other hand, compared to previous representations, hierarchies proved to be effective due to their ability to capture information at different resolution levels. In fact, there is evidence from neuroscience, that information in the visual cortex is structured hierarchically, e.g. for the high-level tasks of recognizing objects (Riesenhuber and Poggio (1999)) or actions (Giese and Poggio (2003)). The structure is usually either defined manually (Zweig and Weinshall (2007)), derived from external lexical resources like WordNet (Rohrbach et al. (2011); Al-Halah and Stiefelhagen (2015)), or based on the similarity in low-level feature space (Salakhutdinov et al. (2010)). However, the manual design of hierarchies is very time consuming, and constructing a hierarchy based on the similarity of low-level data is not favorable for transfer learning, since the feature distributions in source and target may differ significantly. Using lexical databases may sometimes be a good alternative, still in many cases, they are either not complete or not suitable to model all types of information. For example, WordNet does not contain verb-noun combinations to sufficiently describe sports actions, such as *discus-throw* and *javelin-throw*. To overcome the aforementioned problems, we propose a hierarchical model that is learned from the embedded structure in the attribute space and encodes the relative similarities between categories.

Defining a “good” semantic vocabulary is considered an important aspect of the high-level representations. Issues like diversity, coverage and descriptiveness control the quality of the defined semantics. While this aspect is out of the scope of this paper, recent work addressed this problem in the domain of object recognition using web-based weak supervision (Divvala et al. (2014); Berg et al. (2010)), or by leveraging lexical ontologies (Rohrbach et al. (2010)). Incorporating a module to automatically mine good semantics from the source set is quite beneficial for transfer metric learning as it would alleviate the need for human supervision and automate the whole knowledge transfer process.

Transfer Learning Evaluation Protocol
In the evaluation of transfer learning approaches, it is com-

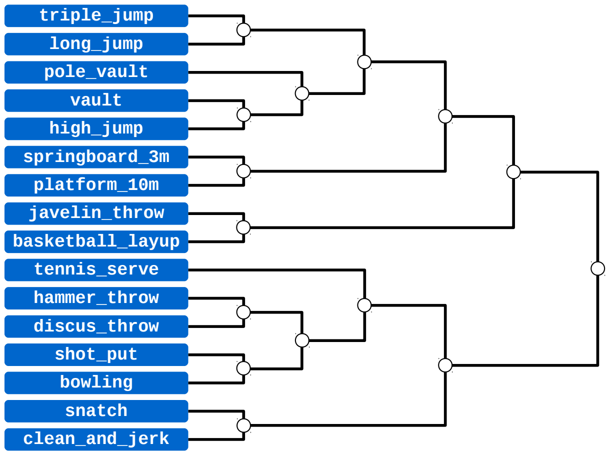


Fig. 2: The learned hierarchical representation of action classes in Olympic Sports where actions are grouped based on their intra-class similarity in the attribute space.

monly assumed that the target set consists of fewer and less diverse classes than the source set (e.g. Lampert et al. (2009); Farhadi et al. (2009); Liu et al. (2011a); Bart and Ullman (2005); Zhang and Yeung (2010)). Even in the large-scale evaluation of Rohrbach et al. (2011), where different transfer approaches for zero-shot object recognition were tested, the source set contained four times more classes than the target set. In this work, we suggest the opposite experimental settings, i.e. the target is much more complex than the source set, resulting in a more challenging evaluation task.

3. Approach

3.1. Semantic Similarity Spaces

Most transfer learning approaches used for object and action recognition are based on low-level features (e.g. Davis et al. (2007); Guillaumin et al. (2009); Zhang and Yeung (2010); Zha et al. (2009)). However, we believe that semantics at different levels of complexity can be a better representation for the transfer of source knowledge 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 more robust to changes in the data distributions between target and source domains. In the following, we will describe the two most common semantic spaces that are used as an intermediate representation, the *attribute similarity space* and the *category similarity space*. In the former, instances are represented by their visual properties, and in the latter, by their resemblance to other previously learned categories. Furthermore, we introduce a third and novel similarity space, the *hierarchical similarity space*. Here, the instances are represented by a hierarchical structure, that captures their visual properties at different resolution levels.

*3.1.1. **Attribute Similarity Space***

Attributes define an intermediate representation between low-level features and high level categories (Lampert et al. (2009); Farhadi et al. (2009)). Semantic attributes describe an entity regarding its visual appearance (e.g. *is-round*), parts (e.g. *has-ears*), and motion patterns (e.g. *forward-motion*). Hence, they can be easily shared across categories and even used to predict unseen classes if the classes can be described in terms of the same vocabulary. In the attribute similarity space $\mathcal{A}$, the different semantic attributes span the bases 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, see Figure 1a. 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 lower the distance between points in $\mathcal{A}$, the more attributes they have in common, and consequently, the more similar they are conceptually. The samples in the d-dimensional feature space $\mathcal{X}^d$ are mapped to space $\mathcal{A}$ using:

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

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

*3.1.2. **Category Similarity Space***

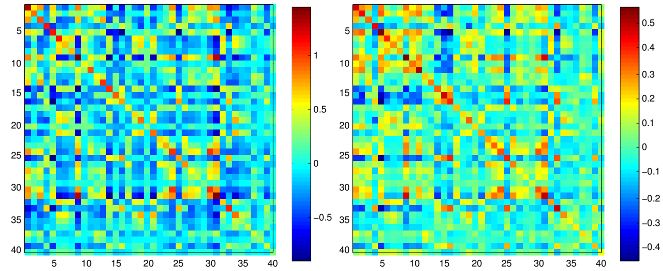
Humans do not only use visual properties to describe entities in their environment, but also inter-class relationships. Consider for example the action class *triple-jump*; it can be described as an action similar to the classes *run* and *jump*. This intra-class similarity pattern is not specific to a certain sample of *triple-jump*, rather it characterizes all samples that belong to this category. In that sense, the category similarity space C provides a meaningful semantic space to compare different actions in terms of their similarity patterns to previously learned categories (Bart and Ullman (2005)). In C, the bases are spanned by the predefined categories, where each axis encodes the resemblance of a sample to a learned category, see Figure 1b. Samples from the feature space are mapped to C using:

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

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

*3.1.3. **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 action categories *walk*, *jump* and *jump-forward*. Since the latter class is partially similar to the former ones, it would be better represented by a super-class consisting of the other two categories, i.e. by *walk-jump*. Then learning the common pattern between these two classes could provide a higher category of semantics that improves the classification performance for *jump-forward*.



(a) (b)

Fig. 3: The correlations of semantic attributes on (a) Olympic Sports and (b) ASLAN (Best seen in color).

Constructing such a hierarchy based on low-level features will not necessarily result in a semantically meaningful structure. Hence, we propose to learn the structure of 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. Thus, assuming that each of the action categories is described with a vector of semantic attributes of length n ($\mathbf{a}^{c_i} = \{a_j\}_1^n$), we can exploit this representation by defining a distance function f to group categories close to each other based on their similarity in the attribute space (Figure 2), i.e.:

$$f : C \times C \rightarrow \mathbb{R} : f(c_i, c_j) = \mathrm{d}(\mathbf{a}^{c_i}, \mathbf{a}^{c_j}), \tag{3}$$

where $\mathrm{d}(\cdot, \cdot)$ is a distance function in the attribute space.

We construct a hierarchical representation by applying an agglomerative hierarchical clustering algorithm over the attribute representation of the classes to get a dendrogram depicting the hierarchical clustering result. The dendrogram is then used to construct the final action 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 our case of using binary attributes to describe the various action classes, we use a hierarchical k-means clustering algorithm with $f(c_i, c_j) =\| \mathbf{a}^{c_i}, \mathbf{a}^{c_j} \|_1$ to capture the similarity in $\mathcal{A}^n$. Then instances $x \in \mathcal{X}^d$ are represented in the hierarchical similarity space $\mathcal{H}$ using:

$$\begin{aligned} f_{\mathcal{H}}(x) \quad &: \quad \mathcal{X}^d \rightarrow \mathcal{H}^k \text{ and} \\ f_{\mathcal{H}}(x) \quad &= \quad [f_{n_1}(x), f_{n_2}(x), \ldots, f_{n_k}(x)]^T, \end{aligned} \tag{4}$$

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 a child-vs-parent manner, i.e. if $\mathrm{pos}(n_i) = \bigcup \mathrm{pos}(n_j)$ is the positive set of node n_i, where $n_j \in \mathrm{child}(n_i)$, then the classifier f_{n_i} is trained on $\mathrm{pos}(n_i)$ as the positive set against $\{\mathrm{pos}(n_p)/\mathrm{pos}(n_i)\}$ as the negative set, where $n_p = \mathrm{parent}(n_i)$.

3.2. Decorrelated Normalized Space

It is important to notice that when the semantic similarity spaces are learned, also the correlations of the semantics are implicitly modeled in theses spaces. Most likely, these correlations are significantly different between the source and the target domain since they arise from the respective semantics distribution in each domain. For example, Figure 3 shows the respective different correlations of semantic attributes in two different data sets. Maintaining such knowledge in the representation when transferring across domains will likely results in a negative transfer effect (Pan and Yang (2010); Torrey and Shavlik (2009)). Therefore, it is quite important to eliminate the correlations learned in the source domain from the semantic spaces in order to restrain the negative transfer.

The decorrelation of the semantic similarity space $\mathcal{S}$ ($\mathcal{S} \in \{\mathcal{A}, C, \mathcal{H}\}$) can be efficiently achieved using the whitening transformation. Such a transformation has been successfully used before for attribute decorrelation (Al-Halah et al. (2014a)) and for removing co-occurrence patterns from the bag-of-words model (Jegou and Chum (2012)), for example. The correlations are modeled by the covariance matrix $\mathbf{\Omega} = \mathbf{Y}\mathbf{Y}^T$ where $\mathbf{Y}$ represents the data matrix from space $\mathcal{S}$. By transforming $\mathbf{\Omega}$ to the identity matrix, $\mathbf{Y}$ is whitened and the data is transformed to a space $\tilde{\mathcal{S}}$ where the bases are decorrelated and given same importance.

The whitening transformation $\mathbf{W}$ of $\mathcal{S}$ is obtained by analyzing the covariance matrix $\mathbf{\Omega}$ such that:

$$\mathbf{W} = \mathbf{V}\mathbf{\Sigma}^{-1/2} \text{ and } \mathbf{\Omega} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T, \tag{5}$$

where $\mathbf{\Sigma}$ is a diagonal matrix having the eigenvalues of $\mathbf{\Omega}$ as its diagonal elements ($\mathbf{\Sigma}_{ii} = \lambda_i$). $\mathbf{V}$ contains in its columns the relevant eigenvectors of the covariance matrix. To have a robust estimation of $\mathbf{W}$, we ignore the eigenvectors in $\mathbf{V}$ that correspond to very small eigenvalues ($\lambda_i < \theta$), i.e.:

$$\hat{\mathbf{W}} = \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{-1/2} \text{ where } \hat{\mathbf{\Sigma}}_{ii} \geq \theta. \tag{6}$$

To have a better assessment of the similarity, we normalize the vectors in the truncated whitened space by their norms. Thus, the samples representation in $\mathcal{S}$ is transformed to the decorrelated normalized space $\mathcal{S}_{dns}$ using:

$$f_{\mathcal{S}_{dns}}(x) = \hat{\mathbf{W}}^T y / \|\hat{\mathbf{W}}^T y\|_2 \text{ where } y = f_{\mathcal{S}}(x) \tag{7}$$

3.3. Similarity Metric Learning

In order to measure similarity between samples in the different semantic spaces, we need to learn an appropriate metric. For that purpose we use the Logistic Discriminant based Metric Learning (LDML) from Guillaumin et al. (2009) to adapt to the positive and negative similarity relations in the target data set.

LDML formulates the metric learning problem 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 the evaluation (Section 4.1).

4. Evaluation

We evaluate our framework using three publicly available data sets:

- **Olympic Sports** (Niebles et al. (2010)), which 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 Liu et al. (2011a), 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*.

- **ASLAN**, which has been recently published by Kliper-Gross et al. (2012), is collected for the main task of comparing actions (similar/not-similar). It has 432 action classes with more than 3600 video samples and each class has on average 8.5 video samples with more than 100 classes having only one sample each.

- **KTH** (Schüldt et al. (2004)), which contains six basic action classes (i.e. *boxing*, *clapping*, *waving*, *jogging*, *running*, and *walking*). In our experiments the classes are described with 10 semantic attributes by Liu et al. (2011a).

In our experimental settings, we use Olympic Sports (or KTH) as source and ASLAN as target data set. This addresses a realistic and very difficult scenario of transfer learning that has been ignored in previous studies as discussed earlier. Collecting and labeling samples for actions is time consuming and expensive. Consequently, the labeled data (source set) tends to be small and simple in terms of diversity and coverage compared to the target. Our evaluation setup tackles this 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) from Laptev et al. (2008) with a vocabulary of size 4000. We use that BoW model to train the different classifiers, presented in Section 3.1, on the training split of Olympic Sports. The features are preprocessed with a power transform (Arandjelovic and Zisserman (2012)) with $\alpha = 0.3$ before training a linear support vector machine. The parameters of the SVM 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.

4.1. Importance of Decorrelated Normalized Space

We first evaluate the impact of the proposed decorrelated normalized space (DNS) on the transfer process effectiveness. We learn the different knowledge representations on Olympic Sports and transfer them to ASLAN where we use the view 1 training/testing split as defined by Kliper-Gross et al. (2012). We test our framework with and without the DNS transformation.

Table 1: The effect of the decorrelated normalized space (DNS) on the performance of popular metric learning methods.

Space / Metric	ITML	LDML	KISSME	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**

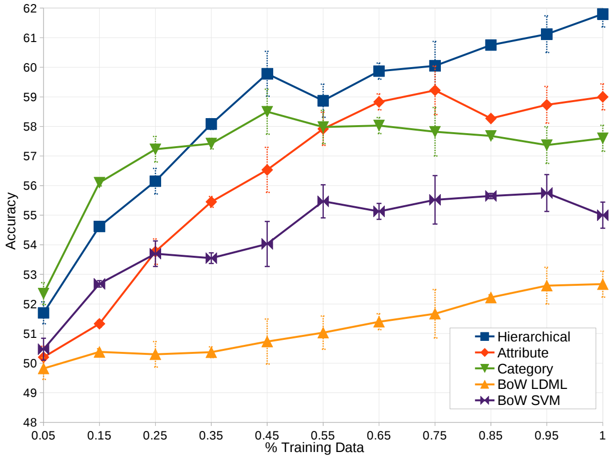


Fig. 4: Overall performance of different semantic similarity spaces regarding various sizes of the target training set. The transferred high-level semantics clearly outperform the low-level representation.

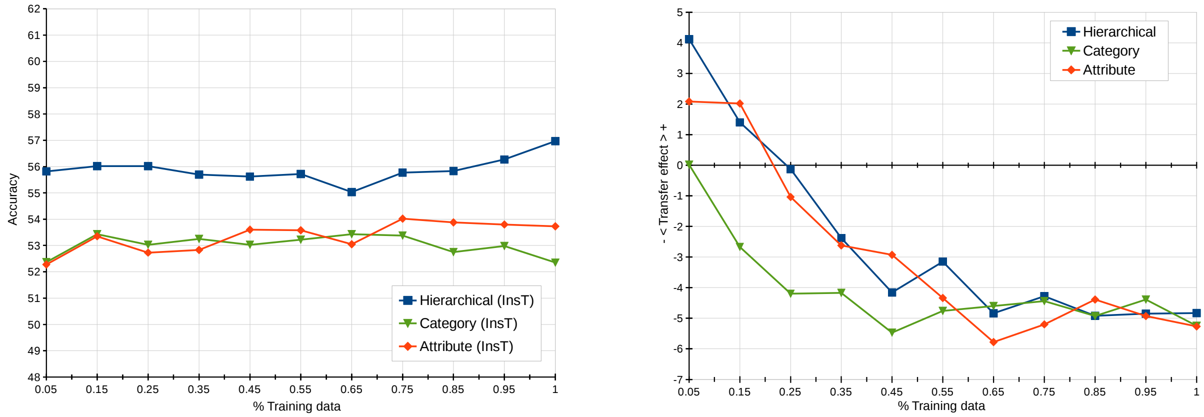
Furthermore, since our framework is not restricted to a specific metric learning approach, we test (along with LDML) two state-of-the-art metric learning methods: ITML (Davis et al. (2007)) and KISSME (Martin et al. (2012)); and two commonly used metrics: the Mahalanobis distance using the inverse of the covariance (Cov^{-1}) and the euclidean distance (L_2). Additionally, we train an SVM on the element wise multiplication of the training pairs $[x_1.{*}x_2]$ as the sixth approach for learning similarities (using the absolute difference $|x_1 - x_2|$ or the concatenation of the previous two produced inferior performance). We use all training pairs available in the target and report the accuracy of the different knowledge representations with and without using DNS.

In Table 1, 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). DNS is quite generic, and it improves the performance of most of the metric learning approaches. Even when using simple metrics like L_2 and Cov^{-1}, DNS helps to learn a better similarity metric.

On the other hand, both the category and hierarchical spaces appears to perform better than the attribute model; and the best performance (61.80%) is obtained by using our hierarchical model with LDML.

4.2. Knowledge Representation Transfer

We test the performance of different semantic spaces compared to the common low-level similarity space. Similar to the previous experiment, the various representations are learned in



(a)

(b)

Fig. 5: The performance of the different knowledge representations when using instance transfer (a) and its negative transfer effect (b).

Olympic Sports and we evaluate using view 1 in ASLAN. However, this time we vary the number of training pairs of similar and dissimilar actions 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 standard error of similarity classification as seen in Figure 4. For the feature space, we report two methods: the first is using LDML after reducing the dimensionality to 128 using principle component analysis since it is intractable to use the full feature vector with LDML (Guillaumin et al. (2009)). In the second we use the full feature vector (4000) and train an SVM on the element wise multiplication similar to Section 4.1.

In compliance with the observations from the previous experiments, we see that the three semantic spaces outperform the low-level feature space (Figure 4). Moreover, 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 outperforms both. This confirms our previous hypothesis on the importance of high-level semantics and their ability to generalize well when transferred to other domains.

Another interesting aspect of the high-level semantics is their scalability. The high-level representation is much more compact than its low-level counterpart. For example, in our case the dimensionality of semantic representations ranges between 16 to 40 while the low-level feature vectors are of 4000 dimensions. Consequently, the semantic representations are more scalable to big data sets since the computation cost of most of the metric learning algorithm is heavily impacted by the representation dimensionality. Moreover, adding new concepts for the attribute and category similarity spaces results in a linear expansion in the dimensionality of the similarity space where only the new concept classifiers need to be trained. Adding a new concept to the hierarchical space is equivalent to inserting a leaf node to a binary tree. It requires the retraining of the ancestors of that leaf node which is of logarithmic complexity in term of the number of nodes in the tree. This cost is much lower than trying to increase the descriptiveness of the low-level features which usually results in much higher computation cost. For instance, adding a new cluster to the bag-of-words requires rerunning the clustering algorithm over all samples again.

4.3. *Instance Transfer*

In this transfer setup, a random group of training pairs from the source (Olympic Sports) are added to the training set in the target (ASLAN). Similar to the previous experiments, we vary the size of the target's training set and report the accuracy. We also analyze the transfer effect (positive or negative) as the difference in performance (measured by accuracy) between using instance transfer and without using it.

We see in Figure 5b that when the target's data is too small (less than 25% of the training pairs), both the hierarchical and attribute spaces take advantage of the additional transferred samples from the source. However, when the size of target training set increases, the transferred instances prevent the metric learning to adapt to the actual data distribution of the target. Hence, it produces a significant negative transfer for all semantic spaces. Nonetheless, the hierarchical representation still maintains higher performance compared to the other alternatives (Figure 5a).

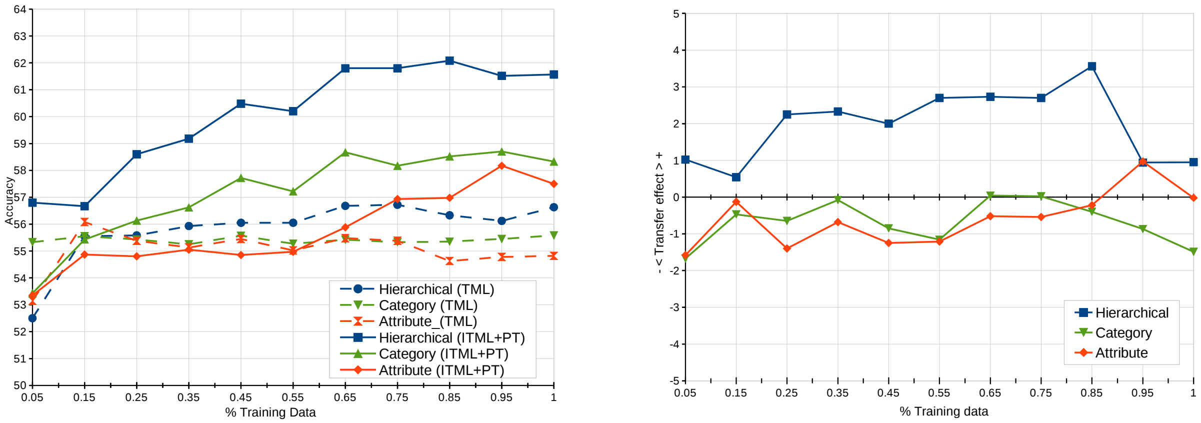
This type of transfer introduces an extreme change in the data distribution of the target training set which is not reflected in the test set, resulting in performance deterioration. It also shows how the target and source sets are different and how challenging are the transfer settings.

4.4. *Parameter Transfer*

In parameter transfer, the parameters learned in the source domain are used to regularize or to aid the learning task in the target. 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) from Davis et al. (2007). The metric learning problem in ITML is defined as:

$$\min_{\mathbf{M}} \ \mathrm{KL}(p(x, \mathbf{M}_0) \parallel p(x, \mathbf{M})), \tag{8}$$

where KL is the Kullback-Leibler divergence between two Gaussian distributions corresponding to a prior metric $\mathbf{M}_0$ and



(a) (b)

Fig. 6: (a) Comparison of the proposed parameter transfer approach (ITML+PT) to the state-of-the-art (TML) and (b) its robustness using the various semantic representations to negative transfer effect.

the learned metric $\mathbf{M}$. Additionally, some constraints on the distances are incorporated in learning the metric:

$$\begin{array}{lll} d_{\mathbf{M}}(\mathbf{v}_i, \mathbf{v}_j) & \leq u & if\ (\mathbf{v}_i, \mathbf{v}_j) \in \mathbf{S} \\ d_{\mathbf{M}}(\mathbf{v}_i, \mathbf{v}_j) & \geqslant l & if\ (\mathbf{v}_i, \mathbf{v}_j) \in \mathbf{D}, \end{array} \tag{9}$$

where u and l are the upper and lower bound of distances between similar ($\mathbf{S}$) and dissimilar ($\mathbf{D}$) pairs respectively.

In Eq. 8, the common assumption is that the data is Gaussian distributed and the prior $\mathbf{M}_0$ is either set to the inverse of the covariance matrix or the identity matrix $\mathbf{I}$ (euclidean metric). In contrast, we suggest to adapt ITML to carry on parameter transfer by setting the prior to be the metric learned in the source data set ($\mathbf{M}_0 = \mathbf{M}_{source}$). In other words, following Eq. 8, 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 (Eq. 9).

We evaluate the parameter transfer setting by learning first the similarity metric for each of the three semantic spaces (Section 3.1) in the source set (Olympic Sports) and transfer that metric using Eq. 8 to the target set (ASLAN). The metric in Olympic Sports is learned by randomly generating 1500 pairs of similar and dissimilar actions in the source, and then using the standard proposed framework to learn the similarity. During testing, we use the same settings as described in Section 4.2.

We compare ITML with the proposed parameter transfer approach (ITML+PT) to state-of-the-art transfer metric learning (TML) from Zhang and Yeung (2010). The parameters for both ITML and TML are set following the recommendations suggested by Davis et al. (2007) and Zhang and Yeung (2010), respectively.

Interestingly, ITML+PT outperforms TML (Figure 6a). TML seems to have a saturated performance after using just 15% of the training set and slightly profits from the different semantic representations. ITML+PT, on the other hand, clearly takes advantage of the characteristics of the different similarity spaces and has a higher initial performance. 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.

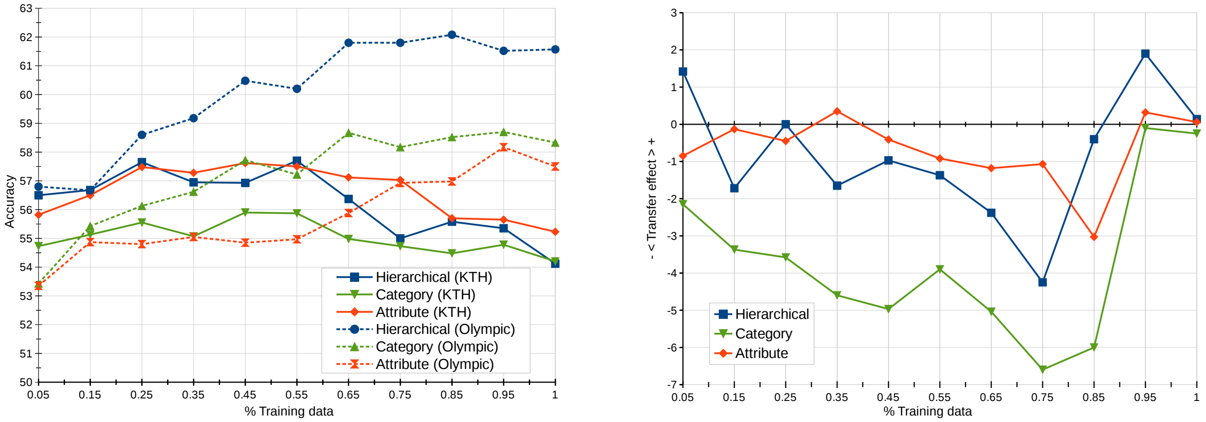
We analyze the transfer effect (as in Section 4.3) as the difference in performance between using the parameter transfer and without (i.e. setting $\mathbf{M}_0 = \mathbf{I}$ in Eq. 8). While the hierarchical representation evidently benefits from parameter transfer, both the attribute and category similarity representations show a negative transfer effect (Figure 6b). As motivated in Section 1, it seems that the robustness of the model against negative transfer is increased when the level of semantic knowledge encoded in it is higher. After all, the learned meta-information (parameters) in source domain can still be true in the target even though they have very different data distributions.

4.5. *Effect of the Source Complexity*

While Olympic Sports contains videos collected from YouTube with a lot of variations (like camera motion, occlusion and varying background), KTH contains only simple motion patterns and is recorded with a uniform background. In this experiment, we test the effect of replacing Olympic Sports with the more simpler KTH data set as the source of the transfer metric learning.

We use a similar setup as in Section 4.4. A similarity matrix ($\mathbf{M}_{source}$) is learned in KTH (the source set) from a set of randomly generated pairs of action samples. Then, $\mathbf{M}_{source}$ is transferred using ITML+PT for metric learning in ASLAN.

Figure 7a shows the performance of the transfer process when using KTH against Olympic Sports as the source set. In general, KTH-based transfer performs worse than the alternative source data set. Another observation is that the KTH-based transfer performance curves do not monotonically increase as its Olympic-based counterparts. The performance of the various KTH-based knowledge representations start to exhibit a drop when the training set in the target gets bigger than 50%. This is expected since KTH contains much less variation in its samples. Hence, it is harder to extract rich semantic representations and learn useful similarity relations. This is evident in Figure 7b, where the difference in performance against using $\mathbf{M}_0 = \mathbf{I}$ (i.e. no parameter transfer) for the KTH-based transfer is shown. Clearly, the similarity relations learned among the classes and attributes in KTH do not generalize well to ASLAN and a significant negative transfer is produced. Nonetheless, when the training set in the target is tiny ($\leq$ 15%) the transferred



(a) (b)

Fig. 7: Performance of parameter transfer when using KTH as the source set (a) instead of Olympic Sports and (b) its negative transfer effect.

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

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

knowledge from the very simple source (KTH) aides the learning process in the target, performing on par with their Olympic Sports counterparts. This suggests, that transferring semantics from simple sources may still be beneficial under harsh transfer setting (i.e. extremely scarce target training data).

4.6. *Full Scale Evaluation*

It is common in transfer learning literature to focus in evaluation only on the case when the training data in target is scarce. However, considering the scenario of a large training set in the target is also beneficial. Evaluating in such settings helps us to put the transfer metric learning method in perspective to standard methods that learn knowledge representation in the target set and have enough information to adapt well to the target data distribution.

For that purpose, we evaluate on ASLAN view 2 which has 6000 pairs of similar and dissimilar actions. We follow the benchmark setup suggested by Kliper-Gross et al. (2012). That is, a 10-fold cross validation is carried out on view 2 and the performance is reported in terms of average accuracy and area under receiver operating characteristic (ROC) curve. For an in-target representation modeling, we compare to the approach of Kliper-Gross et al. (2012). They propose to extract three feature types: HOG, HOF, and HNF (Laptev et al. (2008)); and learn a BoW model of size 5000 for each to represent video samples. They use 12 different similarity metrics to compare actions based on each of these three representations and their combination. We report in Table 2 the results of their best single similarity metric and the results of using the combination of the 12 metrics as stated in Kliper-Gross et al. (2012).

We notice in Table 2 that the transfer metric method performs as well as the methods that are based on 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 only 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. Furthermore, the performance of the different semantic spaces in the transfer metric approach follows the complexity level of semantics encoded in the model. The proposed hierarchical representation is doing best, followed by the category, and attribute spaces.

5. Conclusion and Future Work

We proposed a generic framework for transfer metric learning and showed the importance of knowledge representation on different transfer options. In our experiments, we also demonstrated that high-level semantics have better transfer properties and encode richer transferable knowledge in comparison to low-level features. Furthermore, 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 robust 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, a large-scale evaluation showed impressive results of the transfer approach; the performance is on par with methods that use feature representations learned in the target domain.

So far, we only applied our framework to the Action Similarity task and showed promising results. As future work, we plan to analyze the performance of our approach in the context of other problems, and extend it to overcome challenges posed by the different domains. For instance, in case of recognizing composite activities, that consist of a sequence of multiple actions and human-object interactions, we expect that modeling and transferring relationships between the concepts (e.g. actions and objects) would further boost overall performance. Furthermore, recent advancement in distributional word representation (Mikolov et al. (2013); Pennington et al. (2014)) showed impressive performance in encoding and transferring knowledge for zero-shot learning across data sets for both object (Frome et al. (2013)) and recently action recognition (Xu et al. (2015); Gan et al. (2015)). Exploiting such representation for transfer metric learning could be very beneficial since the representation is learned from large text corpora and requires no human supervision.

Acknowledgment

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

References

Al-Halah, Z., Gehrig, T., Stiefelhagen, R., 2014a. Learning Semantic Attributes via a Common Latent Space, in: VISAPP. doi:10.5220/0004681500480055.

Al-Halah, Z., Rybok, L., Stiefelhagen, R., 2014b. What to Transfer? High-Level Semantics in Transfer Metric Learning for Action Similarity, in: ICPR. doi:10.1109/ICPR.2014.478.

Al-Halah, Z., Stiefelhagen, R., 2015. How to Transfer? Zero-Shot Object Recognition via Hierarchical Transfer of Semantic Attributes, in: WACV. doi:10.1109/WACV.2015.116.

Arandjelovic, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval, in: CVPR.

Bart, E., Ullman, S., 2005. Single-example learning of novel classes using representation by similarity, in: BMVC.

Bellet, A., Habrard, A., Sebban, M., 2013. A Survey on Metric Learning for Feature Vectors and Structured Data. Technical Report.

Berg, T.L., Berg, A.C., Shih, J., 2010. Automatic Attribute Discovery and Characterization from Noisy Web Data, in: ECCV.

Cook, D., Feuz, K.D., Krishnan, N.C., 2013. Transfer Learning for Activity Recognition: A Survey. KAIS 36, 537–556.

Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S., 2007. Information-theoretic metric learning, in: ICML.

Divvala, S.K., Farhadi, A., Guestrin, C., 2014. Learning Everything about Anything: Webly-Supervised Visual Concept Learning, in: CVPR.

Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing Objects by their Attributes, in: CVPR.

Fei-Fei, L., 2006. Knowledge transfer in learning to recognize visual object classes, in: ICDL.

Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model, in: NIPS.

Gan, C., Lin, M., Yang, Y., Zhuang, Y., Hauptmann, A.G., 2015. Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition, in: AAAI.

Giese, M., Poggio, T., 2003. Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience 4, 179–192.

Gong, B., Shi, Y., Sha, F., Grauman, K., 2012. Geodesic Flow Kernel for Unsupervised Domain Adaptation, in: CVPR.

Guillaumin, M., Verbeek, J., Schmid, C., 2009. Is that you? Metric Learning Approaches for Face Identification, in: ICCV.

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

Kliper-Gross, O., Hassner, T., Wolf, L., 2012. The action similarity labeling challenge. T-PAMI .

Lam, A., Roy-Chowdhury, A.K., Shelton, C.R., 2010. Interactive Event Search Through Transfer Learning, in: ACCV.

Lampert, C., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer, in: CVPR.

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies, in: CVPR.

Liu, J., Kuipers, B., Savarese, S., 2011a. Recognizing Human Actions by Attributes, in: CVPR.

Liu, J., Shah, M., Kuipers, B., Savarese, S., 2011b. Cross-View Action Recognition via View Knowledge Transfer, in: CVPR.

Long, M., Ding, G., Wang, J., Sun, J., Guo, Y., Yu, P.S., 2013. Transfer Sparse Coding for Robust Image Representation, in: CVPR.

Martin, K., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H., 2012. Large Scale Metric Learning from Equivalence Constraints, in: CVPR.

Mikolov, T., Corrado, G., Chen, K., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space, in: ICLR.

Nater, F., Tommasi, T., Grabner, H., Van Gool, L., Caputo, B., 2011. Transferring Activities: Updating Human Behavior Analysis, in: ICCV Workshop on Visual Surveillance.

Niebles, J.C., Chen, C.W., Fei-Fei, L., 2010. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, in: ECCV.

Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2011. Domain adaptation via transfer component analysis. NNLS 22, 199–210.

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. TKDE 22, 1345–1359.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe : Global Vectors for Word Representation, in: EMNLP.

Reder, L.M., Klatzky, R., 1994. Transfer: Training for performance, in: Druckman, D., Bjork, R.A. (Eds.), Learning, Remembering, Believing: Enhancing Human Performance. National Academy Press.

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025.

Rohrbach, M., Stark, M., Schiele, B., 2011. Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting, in: CVPR.

Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B., 2010. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer, in: CVPR.

Salakhutdinov, R., Tenenbaum, J., Torralba, A., 2010. Learning to Share Visual Appearance for Multiclass Object Detection, in: CVPR.

Schüldt, C., Laptev, I., Caputo, B., 2004. Recognizing Human Actions: a local SVM Approach, in: ICPR.

Torrey, L., Shavlik, J., 2009. Transfer Learning. Handbook of Research on Machine Learning .

Valiant, L.G., 1984. A theory of the learnable. Communications of the ACM 27, 1134–1142.

Woodworth, R.S., Thorndike, E.L., 1901. The influence of improvement in one mental function upon the efficiency of other functions (I). Psychological Review 8, 247–261.

Xu, X., Hospedales, T., Gong, S., 2015. Semantic embedding space for zero-shot action recognition, in: ICIP.

Zha, Z.J., Mei, T., Wang, M., Wang, Z., Hua, X.S., 2009. Robust Distance Metric Learning with Auxiliary Knowledge, in: IJCAI.

Zhang, Y., Yeung, D.Y., 2010. Transfer metric learning by learning task relationships, in: ACM SIGKDD - KDD.

Zweig, A., Weinshall, D., 2007. Exploiting Object Hierarchy: Combining Models from Different Category Levels, in: ICCV.