Cross-domain Learning in Deep HAR Models via Natural Language Processing on Action Labels

Konstantinos Bacharidis[1,2] and Antonis Argyros[1,2]

[1] Computer Science Department, University of Crete, Greece
[2] Foundation for Research and Technology - Hellas, Heraklion, Greece
{kbach,argyros}@ics.forth.gr

Abstract. Nowadays, deep learning approaches lead the state-of-the-art scores in human activity recognition (HAR). However, the supervised nature of these approaches still relies heavily on the size and the quality of the available training datasets. The complexity of activities of existing HAR video datasets ranges from simple coarse actions, such as sitting, to complex activities, consisting of multiple actions with subtle variations in appearance and execution. For the latter, the available datasets rarely contain adequate data samples. In this paper, we propose an approach to exploit the action-related information in action label sentences to combine HAR datasets that share a sufficient amount of actions with high linguistic similarity in their labels. We evaluate the effect of inter- and intra-dataset label linguistic similarity rate in the process of a cross-dataset knowledge distillation. In addition, we propose a deep neural network design that enables joint learning and leverages, for each dataset, the additional training data from the other dataset, for actions with high linguistic similarity. Finally, in a series of quantitative and qualitative experiments, we show that our approach improves the performance for both datasets, compared to a single dataset learning scheme.

Keywords: Human action recognition · Natural language processing · Deep learning · Video understanding.

1 Introduction

In recent years deep learning has become the dominant learning direction in several research fields, including computer vision. Human activity recognition (HAR) is one of its challenging sub-fields, with a wide range of applications from Human-Robot Collaboration (HRC) and assistive technologies for daily living, to surveillance and entertainment. Deep learning models have dominated the field due to their high representational power, long-range temporal modelling capacity, as well as their end-to-end training capabilities. The majority of these models rely on a supervised learning process, with the most powerful ones requiring large-scale datasets with diverse video content and action/activity sets, especially for layer-related hard optimization cases, such as 3D convolutional filter-based ones. However, the number of publicly available large-scale HAR

datasets is rather small. The most common workaround to improve performance and generalization on small-scale datasets is to exploit a model that has been trained on large-scale image or video recognition datasets, such as ImageNet [9] or Kinetics [4], as a generic feature extractor and only train a shallow temporal model on the target dataset [14], or fine-tune the entire spatio-temporal model [26,11], a concept known as *transfer learning*.

Another direction is to consider action category commonalities between dataset pairs, and apply a joint learning scheme (*multi-task learning*) for the two action domains [21], leveraging of additional data for the class set that lies in the shared label space, referred as *supervised Domain Adaptation (DA)* [30]. The evaluation of the contribution of this learning tactic is carried out in carefully selected dataset pairs that fulfill the criteria of having a sufficient number of common action classes and similar motion and appearance characteristics, in order to constrain the distribution gap due to the *domain shift*. In the existing literature, there exists only a limited number of such dataset pairs, which are defined via manual evaluation of the aforementioned attributes [7,5]. Under this premise, the development of a generalized framework for automatically evaluating the potential compatibility of two or more datasets, is an interesting but still unaddressed research direction. Our work is an attempt to tackle this problem, with a flexible and interpretive domain adaptation-oriented dataset association process based on label linguistic similarities for the considered datasets.

2 Related Work

Cross-Domain learning in action recognition: aims at reducing the distribution gap between the feature spaces of the considered domains through joint modelling. To achieve this, existing works have incorporated feature distribution similarity measures, such as the Kullback–Leibler (KL) divergence, and the Maximum Mean Discrepancy (MMD), along with the task of image [17,2], video classification [31,5]. Expanding on the task of action recognition, a set of deep learning works, instead of only relying on distribution similarity error metrics, attempt to reduce the domain gap at feature level, by introducing domain alignment layers that consider batch-level statistics and cross-domain batch contamination strategies [3,21] in their designs of a cross-dataset HAR learning deep model, which operates on the concatenated label set of the datasets.

Dataset association: has been considered in numerous works, as a means to increase the generalization of models, expand the supported label space, and handle imbalanced datasets [22,27]. In the contexts of video cross-domain learning and DA, existing works have combined dataset pairs with a range of approaches. These approaches include simple strategies, such as formulating a new dataset comprised of the union of the label sets [21] or re-annotating the labels of the second dataset following the annotation protocol of the first [15]. Delving into the task of DA, a set of works considers only common actions between datasets to define the basis in which the shared latent subspace is defined [12,25]. This set of common action classes can be further expanded by grouping semantically similar action labels, considering notions such as word semantic similarity and

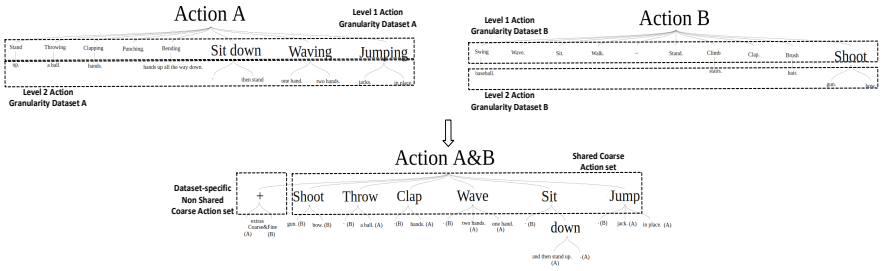

Fig. 1. Hierarchical action label decomposition in coarse, fine action levels via verb-POS analysis. Datasets: *MHAD (A), J-HMDB (B)*

lexical hierarchy. These linguistic associations are usually exploited indirectly via the inherent linguistic knowledge of the annotators, either in the form of direct relabelling of the source dataset to the target [5], or, to provide annotations regarding linguistic and semantic relations between the two label sets [28]. The advantage of the second approach is that these intermediate annotations allow to further analyze the characteristics of the datasets, to compute the general relevance score between the datasets, as well as to generation of a range of dataset label fusions, by considering stronger or weaker label associations.

Different from these works, our work does not utilize annotators for the derivation of the linguistic similarity between the labels, but instead exploits the word semantic similarities and relations from large lexical databases, such as WordNet [19], to define and control the strength of the label associations. In addition, by exploiting dataset relevance statistics, in a similar basis with the work of Yoshikawa et al [28], we are able to evaluate the resulting association. Finally, we investigate the impact of the cross-dataset linguistic similarity rate requirements and single dataset inter-class linguistic label sentence correlation rate on the potential performance gain in a HAR deep model design that exploits the joint label space in a multi-task learning scheme. The core design direction for this model follows the principles of HAR-oriented DA models, and can be related with the work of Bousmalis et al [2], in that we also follow a combined dataset-wise (private) and shared subspace learning scheme. In addition, compared to [2] despite mitigating the problem of cross-domain knowledge transfer to the task of action recognition, our model design aims to learn discrete representations for both datasets and their respective action sets (*multi-task learning*).

3 Proposed Method

The proposed method provides a framework for the relation of video HAR datasets based on the linguistic similarities of their label sentences. Our work shows that such associations can be exploited in a dual-dataset learning scheme applicable to any deep HAR architecture with minor modifications. We argue that such learning schemes and architectures access a richer training sample pool for action classes that share the same or semantically similar linguistic defini-

tions. Our experiments show that as this sample pool size increases, the model's representational strength is enriched, leading to better action discrimination.

3.1 Dataset label association via NLP

The proposed method pipeline operates as follows. First, we present the NLP tools utilized in the computation of label linguistic similarity, and, a label decomposition process that provides an interpretive and precise definition of the linguistic association between action labels. This process transforms the label set of each dataset into two action granularity-based label sets, (a) a *coarse-grained* action set, consisting of simple verb-based labels that denote the common action motif between a set of associated actions (for the actions *get the cup* and *get the bottle*, coarse label is the verb *get*), and, (b) a *fine-grained* action set, with the initial labels enriched with coarse-grained membership information. Subsequently, we present a process to define association rules between a dataset pair and highlight key elements and assumptions of this dataset relation process.

Dataset-wise label association and hierarchical decomposition: In our recent work [1] we presented an NLP-assisted label sentence analysis approach to define a two-level action tree hierarchy from a given set of action labels, either focusing on a specific part-of-speech (POS) or by exploring the semantic relations between the entire label (via word-ordering & semantic content similarities) relying on the work of Yuhua Li et al. [16]. In this work, we also follow a verb-POS action label direction to group semantically similar labels based on verb commonalities, or high verb semantic content similarity.

For the latter case, we evaluate the semantic relation between the verbs of the label sentences based on two metrics. The first metric expresses the semantic relation as defined within the WordNet [19] semantic knowledge base. We define the verb semantic similarity rate between a label pair by thresholding the normalized (to [0,1]) length of the shortest path between the word (verb) nodes relatively to the common word-ancestor node, as defined in WordNet, following the direction of Redmon and Farhadi [22]. The second metric follows a more simplified direction and directly compares the word embeddings of the two words (verbs), generated via the Word2Vec [18] embedding model, using the cosine similarity metric. We found that combining these metrics best expresses the relation between the label sentences in terms of verb semantic content similarity.

Given the detected label associations we can define a two-level action hierarchy based on the verb semantic similarities between the action classes. The first action tree level, consists of a set of coarse action classes, defined by the *shared* verbs[3], indicating the presence of a common coarse motion pattern between the related actions. The second level contains the fine-grained action classes, belonging to the dataset's original set, enriched with info regarding the *coarse* class to which each fine-grained label has been clustered. We should mention that a

[3] For associated labels with different verbs, with high semantic similarity, the verb of 1st label is used as a coarse label.

more complex hierarchy could surface more informative clues[4], however more complex semantic relation trees are scheduled to be explored in the future.

Inter-dataset label association: In a similar fashion, to associate a dataset pair, we utilize NLP to identify action labels that are common or exhibit high semantic similarity, focusing only on the verb POS sets (coarse classes), and fuse the two action trees into a shared, two-level action tree hierarchy. The first level now contains a set of coarse action classes that correspond to the verb-POS elements that are shared between the class sets of the dataset pair, indicating a similar coarse action primitive, as well as the remaining unique coarse classes of both datasets. The second level consists of the fine-grained classes for which a coarser action class was defined. Figure 1 shows a simplified illustration.

In more detail, for a pair of datasets A, B, with isolated verb label sets noted as T_A and T_B, we define the shared coarse action label set C, with the verbs-POS of the labels k, l in A, B, whose verbs are the same, $T_A \cap T_B$, or, (a) the relative path length in WordTree between $verb_k \in T_A, verb_l \in T_B \leq 0.5$, and (b) the cosine similarity between $verb_k \in T_A, verb_l \in T_B \geq 0.9$. The gains for each dataset from this formulation depend on the portion of action labels for each dataset that are shared. A simple, intuitive criterion to define the dataset label set fusion compatibility, is to set thresholds on the minimum portion of labels of each dataset that needs to be included in the shared, coarse label set. Based on this, we can define the label set compatibility for the dataset pair as follows:

Criterion for assessing the dataset label set compatibility: *$|C \cap T_A| \geq t_1|T_A|$ and $|C \cap T_B| \geq t_2|T_B|$ conditioned that $\frac{t_1+t_2}{2} \geq t_3$, with $t_1, t_2, t_3 \in (0, 1]$.*

The parameters t_1, t_2, t_3 determine the required degree of similarity between the two datasets in order to consider the content of their action sets as correlated. Thresholds t_1, t_2, express the portion of the dataset's class set that is encapsulated in the generated *coarse* action class set C. The degree of the overall dataset pair similarity rate is expressed with t_3. The higher the t_3 value, the larger becomes the requirement for the datasets to exhibit higher label semantic associations. With that in mind, we can define levels for the dataset association power (*low, partial, high*) by setting dataset-appropriate values for t_3. For this purpose in our experiments we evaluated the aspect of inter-dataset compatibility by defining the dataset association levels, (a) $t_3 < 0.3$ - *low*, (b) $0.3 < t_3 < 0.6$ - *partial*, and, (c) $0.6 < t_3 < 0.9$ - *high*, with $t_3 = 1$ signifying full association.

The importance of intra-dataset label similarity: The information gain from the fusion of two datasets will be higher as the amount of associated classes increases. A factor that affects the gain is the dataset-wise intra-class label similarity. Ideally, a high label relation threshold (high cosine similarity, short-length paths between words in WordTree) guarantees that only labels with close semantic contents are associated, and exploit the coarser representation knowledge that is acquired from this learning scheme. However, it is interesting to examine the

[4] For example, we could add a level that defines associations based on nouns, referring to the presence of common objects in different actions.

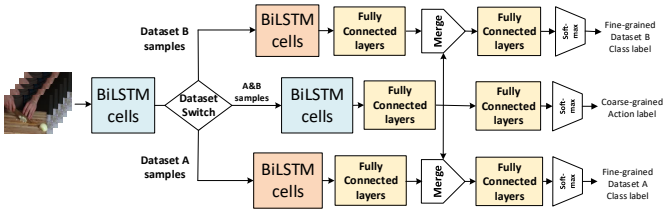

Fig. 2. Baseline BiLSTM DNN for dual-dataset learning. Batch consists of both dataset samples. Each sample contains also a scalar $\in [0, 1]$, indicating dataset membership.

effect of subtle linguistic relations between labels that have been included in the shared set, and the ones that were not. To express this in *set theory*, for the two datasets A and B, and their shared action set C, we define the relative complement of B in A as $A_D : (T_A - T_B)$, and the one of A in B as $B_D : (T_B - T_A)$. Our goal is to assess the performance of the dataset association learning scheme based on the degree of the lexical similarity between the labels in A_D and the ones in C, and, in similar fashion, for B_D and C. In a similar factor assessment direction to the one described for the dataset fusing compatibility, in our experiments we examine the effect of the linguistic similarity rate between the intersection C, and non-intersection, A_D (or B_D) sets, under the same three association levels (*no, partial, high*). In addition, since it is difficult to find different dataset pairs that satisfy these conditions, we design a simple algorithm which, given the requested association condition, splits the MPII Cooking Activities [23], into two subsets whose label sets satisfy the requirements. Details in the next section.

3.2 Dual-dataset learning deep architecture

We now present design directions, applicable to the majority of HAR DNNs, that allow the utilization of the dataset association scheme in a dual-dataset learning format, improving the model's performance on one or both datasets.

The simplest HAR DNN design that allows the support of a dual-dataset learning functionality is to merge the datasets into a new expanded action set, $A \cup B$, and classify an input sequence to the unified action label set. In this work we propose a DNN structure, that mimics the hierarchical action decomposition and dataset relation scheme that we defined earlier. It is a triple-branch DNN design (fig 2), consisting of two distinct sub-nets assigned to model each dataset and an additional sub-net that handles the spatio-temporal modelling of the shared coarser actions. Moreover, skip connections introduce the learned coarse-grained representation as complementary information in the fine-grained sub-nets, guiding them to learn representations towards finer action details.

Regarding the objective function, the network learns a shared representation of two different distributions, thus, we need to evaluate the learned representation for the shared coarse action labels. For this we follow the guidelines of cross-domain learning approaches and use the Maximum Mean Discrepancy (MMD) loss [10] to compute the marginal distribution between the domain distributions. The loss function to be minimized is defined as:

$$L = L_{task} + L_{MMD}(Gen, Fine_A) + L_{MMD}(Gen, Fine_B), \quad (1)$$

where L_{task} refers to the classification problem between the coarse and fine action classes, and L_{MMD} refers to the MMD domain distribution distance loss. In detail, the classification loss is defined as the cross-entropy loss for the two action granularities (coarse, fine):

$$L_{task} = -\sum_{k=0}^{K} T_k^{gen} log\left(Y_k^{gen}\right) - \sum_{i=0}^{1}\sum_{j=0}^{J} w_{i,j} T_{i,j}^{fine} log\left(Y_{i,j}^{fine}\right), \tag{2}$$

with (a) the w_i vector denoting the dependencies between the fine-grained action classes of the dataset i (details in [1]), T^{gen} denoting the ground-truth labels for the joint coarse-grained action set, (c) T_i^{fine} being the ground-truth fine-grained labels for dataset i, and, (d) (Y^{gen}, Y_i^{fine}) being the estimated action classes for the coarse- and the fine-grained action sets for dataset i.

The L_{MMD} loss, is actually the summation of two MMD losses, between the learned shared distribution and each dataset-specific learned distribution. Intuitively, regarding the two design directions, the first is simpler to define and learns a mapping from both input domains to the distinct concatenated output label set. However, HAR model of this design can be harder to train. The reason can be thought as a potential combination of, a) model capacity inadequacy due to the fine-grained label space significant expansion, and, b) label cases with similar characteristics combined with data scarcity, resulting in weaker representations for each class that easily lead to mis-classifications.

3.3 Factors that affect learning

The performance of cross-domain and dataset fusion learning such as the one we propose, is affected by a number of factors. The most important one that affects the efficiency of learning in HAR datasets is the differentiation in the dataset characteristics, such as whether the actions are performed in a constrained or unconstrained environment, under a fixed or with multiple viewing angles, the presence of moving objects in the background etc. In HAR cross-domain learning and domain-adaptation setups, the examined datasets share similar action characteristics and are defined under more controlled conditions, such as environments with static scenes with minimal background motions and noise. This allows for the impact on the representation difference to be smaller since the appearance feature manifold is more constrained. To further restraint the effect of such elements, in our experiments we limit the processing area in the actor's region, removing any background information that may induce a negative affect.

A consequence of the aforementioned domain-related differences between the datasets is the distance between the learned representational sub-space in the feature manifold to which the action set of each dataset is mapped to. Ideally, when working with an action set consisting of the union of the label sets, for the cases of actions that are shared, or associated via linguistic similarities, we expect the learned representations to be mapped closely in feature space. However, in the appearance domain (RGB), variations in the background or in the actor/object characteristics can expand the feature representation subspace of

each action and increase the representation gap between actions with similar coarse motion motifs (take a bottle and grab a glass). To constrain the representation gap for such cases we can work with high-level representation spaces, such as optical flow (OF) or pose-based feature representations. In our experiments we also follow this direction by utilizing OF data of, (a) *the entire scene*, (b) *human body part regions*.

4 Experimental Setup

We evaluate the benefits and constraints of the proposed learning and DNN design scheme on three known HAR datasets of ranging action complexity. The first dataset pair consists of the Berkeley's MHAD [20] (11-classes) and J-HMDB [13] (21-classes) datasets. The specific dataset pair shares a number of six coarser classes. The coarser action set for this dataset pair consists of (a) the common coarser classes for both datasets, (b) the remainder of the coarser action classes for the dataset A (MHAD), and, (c) the remainder of the coarser action classes for the dataset B (J-HMDB). A simplified illustration is shown in Figure 1.

The third dataset that has been explored is Max Planck's Cooking dataset (MPII Cooking Activities [23]), which is used to better understand the significance and impact of the similarity rate on the proposed learning scheme. Specifically, it's action label size and complexity as well as the high inter-class similarity (appearance&motion characteristics) between its action label set makes it ideal to serve as the experimental basis for evaluating the inter- and intra-dataset cases, presented in Section 3.1. To adjust MPII Cooking to this format, we designed a simple algorithmic process that splits the dataset into two subsets that satisfy the specifications of different scenarios of inter- and intra-dataset label linguistic similarity. Details are presented in the next subsection.

For the reported scores, for MPII and J-HMDB, we report the accuracy score on split-1, whereas for MHAD, we follow the provided train/test scheme. Regarding input sources, we focus on the OF domain, and consider two feature design strategies, (a) OF estimates on the actor's region, and, (b) OF estimates on distinct body-parts of the actor. OF data were generated with TV-L1 [29].

4.1 Inter- and intra-dataset evaluation

To evaluate the notions in 3.1, instead of searching for dataset pairs that satisfy the inter- and intra-dataset similarity cases, we manually construct them. For this, we designed a simple algorithmic pipeline that splits MPII Cooking Activities into two subsets $MPII_A$, $MPII_B$, that satisfy a specified configuration for inter- and intra-dataset label linguistic similarity.

To decouple the inter- and intra-dataset similarity factors and assess their impact, the algorithmic process[5] contains two functionality sets:
Inter-dataset: generate random splits of the dataset class set into two subsets, under the condition that the similarity rate between class sets of the two subsets satisfies the required threshold, t_3. To evaluate the satisfaction of the requested

[5] The process utilizes the label set, and, the respective word embeddings.

inter-dataset similarity, we estimate the inter-dataset similarity score. For this, we identify the verb-POS of the labels that have been assigned to each subset and compare them using the metrics presented in Section 3.1. The achieved score is evaluated based on t_3. If the threshold is not satisfied the process is repeated.

Intra-dataset: To evaluate intra-dataset similarity for each of the possible similarity rate scenarios, the initial step of the dataset splitting algorithm is to define an intersecting class set $MPII_C$, and then proceed to gradually add the non-intersecting classes to each subset, checking after each insertion the satisfaction of the conditions of each case. This format allows for all generated splits for each condition to share the same common coarse action set, in order to exclude the impact of this factor from the assessment.

The examined association scenarios for the non-intersecting subsets of $MPII_A$, $MPII_B$, noted as $MPII_{A_D}$, $MPII_{B_D}$, with the intersection $MPII_C$, are:
(1) $MPII_{A_D}$, $MPII_{B_D}$ with a relative large portion of labels with high similarity with the ones in $MPII_C$,
(2) $MPII_{A_D}$, $MPII_{B_D}$ with a relative small portion of labels with high similarity with the ones in $MPII_C$,
(3) $MPII_{A_D}$ with a large portion of labels with high similarity with the ones in $MPII_C$, and, $MPII_{B_D}$ a low,
(4) $MPII_{B_D}$ with a large portion of labels with high similarity with the ones in $MPII_C$, and, $MPII_{A_D}$ a low.

In detail, the process begins with the construction of the label self-similarity matrix (LSM), by computing the pairwise cosine similarity of their respective word-embeddings. Based on the LSM scores, we select the N most similar label pairs, and use them as the basis for the intersection label set, $MPII_C$, assigning from each pair, $label_i$ in subset $MPII_A$, and, $label_j$ in $MPII_B$. The rest of the labels, $MPII - MPII_C$ serve as the label pool to construct $MPII_{A_D}, MPII_{B_D}$. This process involves first clustering these labels, using k-means, based on the linguistic similarity of their verb-POS with the verb-POS of the labels in $MPII_C$, which allows the detection of the labels with the most impact on the intra-dataset similarity scores, $\mathbf{Sim}(MPII_{A_D}, MPII_C)$, $\mathbf{Sim}(MPII_{B_D}, MPII_C)$. For each label in each cluster, we use LSM to find each most similar label in the same cluster, and, in $MPII_C$. We assign each of the two labels to the subset, whose label in $MPII_C$ exhibits the highest similarity with it. After all non-intersecting labels have been assigned to one of the subsets, we compute the intra-dataset similarity scores to evaluate their satisfaction. If the requested thresholds are unsatisfied, we randomly select one label from the clusters with the highest dissimilarity and assign the label to the opposite subset, and, recompute the similarity rates. The process repeats until the goal constraints are satisfied.

4.2 Feature extraction

For optical flow (OF), 16-OF frame sequences were used for the I3D network. Contrary, for the case of the Bi-LSTM based architecture, the OF frames were fed to VGG-16 [24]. We then extracted 2D feature maps from the last 2D layer, resulting in a frame-wise feature tensor of 7-by-7-by-512 for the sequence. For the

Table 1. Performance difference between a single dataset (NM), and, a dual dataset (M) DNN designs. Inputs are OF frame sequences. For MPII, splits are A-31 classes, B-33 classes, with intersection similarity rate of 0.38, leading to 11 coarse classes.

Architecture	**Datasets** Acc.%, **Input: OF**			
Design	*MHAD/JHMDB*		$MPII_A/MPII_B$	
NM-lstm	60.18	38.75	28.17	29.65
M-lstm	**63.59**	**41.87**	**36.45**	**29.74**
NM-I3D [4]	86.37	49.89	47.05	48.33
M-I3D	**90.67**	49.58	46.62	**49.83**

second input modality, we follow the work of Chéron et al [6], to generate frame-wise CNN-based features for the actor's right hand, left hand, upper body, full body and full image regions, utilizing the positions of body joints. This results in frame-wise 5×4096 feature maps. The final descriptor formulation stage of PCNN [6] involves a feature map aggregation scheme, that defines a *spatial descriptor* for each part by computing minimum and maximum values for this part following a *max* and *min* pooling scheme, leading to a 1×512 feature vector for each part per frame, and finally concatenating the resulting body part *spatial descriptors*. In this work we consider motion attributes by using OF as input.

4.3 Temporal modelling architectures

For the evaluation of the proposed DNN directions, we compared baseline single-dataset architectures to their modified proposed dual-dataset versions.

Baseline BiLSTM DNN & modification: we design a two-layer BiLSTM net with three Fully-Connected (FC) top layers, with activation functions, Leaky ReLU x2 and soft-max for classification. Inputs are frame-wise deep embeddings. To support dual-dataset learning, the modifications involve the use of a BiLSTM layer as a shared temporal modelling layer between the datasets, followed by decoupling into three sub-nets tasked with representation learning for datasets A, B, and, *set C of coarse classes*. In detail, the coarse-level sub-net consists of a BiLSTM layer followed by a two-level FC layer set, with Leaky ReLU and soft-max. This sub-net generates probability distribution estimates for coarse-grained classes. Contrary, the fine-grained sub-nets consist of a BiLSTM layer followed by a three-FC layer set, with the first two using Leaky ReLU and dropout, and the last a soft-max. The second FC layer input is the concatenation of the feature maps of the first FC layers of the coarse and the dataset-specific sub-net.

I3D [4] & modification: We maintain the original design up until the last receptive field up-sampling layer-block, using the pre-trained weights on ImageNet [8] and Kinetics [4], and fine-tune the last layers on the new datasets. The design modifications to support the dual-dataset learning scheme, follow the same coarse- and fine-grained sub-network structural principles as previous with the difference of replacing BiLSTM with Conv3D layers.

Table 2. Action recognition performance for the MHAD, JHMDB and MPII datasets between a single dataset (NM), and, a dual dataset (M) DNN designs, d refers to the usage of the MMD loss besides cross-entropy. Input source pose OF features [6].

Architecture	**Datasets** Acc.%, **Input: Body-part OF**			
Design	*MHAD/JHMDB*		$MPII_A/MPII_B$	
NM-lstm	75.18	42.28	32.48	38.33
M-lstm	70.89	47.43	30.39	31.54
M-lstm$_d$	**80.31**	**55.29**	**36.02**	**39.14**

Table 3. Inter-dataset similarity threshold and accuracy. Random split of MPII Cooking under a inter-dataset similarity requirement t_3. At each scenario we generate a new splitting of MPII into $MPII_A$ and $MPII_B$ datasets, C contains $MPII_A \cap MPII_B$.

Threshold	**MPII** Acc.%, **Input: Body-Part OF**		
Value	***Subsets (A/B)***	***C***	***Acc. %***
NM-lstm	37/27	-	**25.50**/30.00
$t_3 < 0.3$, M-lstm$_d$	37/27 (*0.2*)	14	20.62/**30.87**
NM-lstm	31/33	-	32.48/38.33
$t_3 \in$(0.3,0.6),M-lstm$_d$	31/33 (*0.38*)	11	**36.02**/**39.14**
NM-lstm	53/11	-	21.04/48.98
$t_3 > 0.6$, M-lstm$_d$	53/11 (*0.72*)	10	**23.84**/**55.37**

5 Experimental Results

The first set of experiments, shown in Tables 1 and 2, illustrate the contribution of a dual-dataset learning strategy, relying on the label-centered linguistic fusion and action decomposition methodology. We can observe that for both modalities and architecture variations there is a clear benefit, with improvements in accuracy reaching up to 9%. An additional observation is that the BiLSTM-based DNN appears to benefit the most, with improvements being observed in both datasets and modalities. Contrary, the proposed design scheme in an I3D-based model, appears to assist recognition on the small-sized subsets, following the observed learning trend reported in existing dual-dataset learning works [21]. It is noted that in this experimental setup, MHAD has 9 training samples/class (a single view was used), compared to J-HMDB that has around 3-4 times more samples/class. For MPII Cooking, for the specific split, $MPII_B$ has on average 44 samples/class, as opposed to the 47 of $MPII_A$. We aim to publicly release the MPII splits created for intra-dataset evaluation.

MMD loss contribution: Table 2 presents the contribution of the distribution adaptation part of the objective function. We observe that for the body part OF modality the inclusion of this term is crucial for the success of the proposed method, improving recognition accuracy on both datasets.

Table 4. Dataset-wise intra-class linguistic similarity impact. Random split of MPII Cooking with t_{Ssim}=0,34. A, B refer to $MPII_A, MPII_B$, C to $MPII_A \cap MPII_B$.

Threshold	**MPII** Acc.%, **Input: Body-part OF**	
Value	***# classes of (A / B / C)***	***Acc.***
A / B	7/57	37.30/21.01%
Intra-Case 1	7 (*0.32*)/ 57 (*0.31*)/ 4	**44.69/23.15**%
A / B	31/33	21.39/23.72%
Intra-Case 2	31(*0.54*)/ 33 (*0.33*)/ 4	**25.83/30.51**%
A / B	29/35	23.63/25.24%
Intra-Case 3	29(*0.52*)/ 35(*0.38*)/ 4	**30.09/29.32**%
A / B	10/54	43.38/20.12%
Intra-Case 4	10 (*0.33*)/ 54 (*0.46*)/ 4	**55.18/24.89**%

Inter-dataset label similarity rate: In Table 3 we present our findings on the role of the inter-dataset label similarity rate on the proposed learning strategy effectiveness. The results on 3 split versions of the MPII Cooking that satisfy each case (*low, partial, high* relation), show that for the proposed method to be beneficial, the pair has to show partial to high label set linguistic association.

Intra-dataset label similarity rate: In Table 4 we present our findings on the role of the intra-dataset label similarity rate. The obtained results for the 4 identified scenarios (see Section 4.1), indicate that the presence of subtle linguistic similarities between the labels in the intersecting and non-intersecting subsets of a dataset, appear to affect the contribution of the proposed dataset fusion and joint learning scheme. This can be observed from the fact that the inclusion of new labels (in the smaller dataset A), that have high similarity with the labels in the intersecting subset, leads to a decrease in the recognition accuracy.

6 Conclusions and Discussion

We proposed an approach to fuse HAR datasets pairs by exploiting NLP to identify linguistic similarities on the label sets. To exploit such associations, we designed a DNN to allow joint dataset learning, leveraging the dataset association knowledge under a multi-task learning scheme. We evaluated parameters that control its effectiveness like the intra-dataset label similarity. Our method positively affects the performance of HAR DNNs, however its effectiveness requires careful consideration of dataset characteristics and label linguistic similarity.

An aspect of the method open for discussion is the context information locality in Word2Vec's embeddings and the fact that WordTree represents general notions of word semantics. As such, they do not encode semantic relations between a word and other parts-of-speech that co-exist in a sentence. Word2Vec relies on local statistics, incorporates the local context information of the neighboring words to the target word, defined within the corpus. This can lead to

semantic context ambiguities, with words associated to different semantic interpretations. In simpler action datasets, this is not an issue as the label sentence length and semantic context is constrained and simplified. However, for fine-grained datasets larger sentences and multiple verbs/nouns are encountered. Thus, a global word context relationship will lead to more informative embeddings. Non-local embedding methods or DNNs with text sequential ordering and long-range dependency modelling mechanisms will be ideal for label similarity evaluation in such datasets. We aim to explore such methods to enrich the semantic context our method considers.

Acknowledgments

This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for H.F.R.I Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, number 91. The authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU.

References

1. Bacharidis, K., Argyros, A.: Improving deep learning approaches for human activity recognition based on natural language processing of action labels. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020)
2. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29 (2016)
3. Carlucci, F.M., Porzi, L., Caputo, B., Ricci, E., Bulo, S.R.: Autodial: Automatic domain alignment layers. In: 2017 IEEE international conference on computer vision (ICCV). pp. 5077–5085. IEEE (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308. IEEE (2017)
5. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6321–6330 (2019)
6. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition. In: ICCV. pp. 3218–3226. IEEE (2015)
7. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. IEEE (2009)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**(25), 723–773 (2012)
11. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 3154–3160 (2017)

12. Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K.: Deep domain adaptation in action space. In: BMVC. vol. 2, p. 5 (2018)
13. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. pp. 3192–3199. IEEE (2013)
14. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2014)
15. Li, A., Thotakuri, M., Ross, D.A., Carreira, J., Vostrikov, A., Zisserman, A.: The ava-kinetics localized human actions video dataset. arXiv preprint arXiv:2005.00214 (2020)
16. Li, Y., McLean, D., Bandar, Z.A., Crockett, K., et al.: Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge & Data Engineering (8), 1138–1150 (2006)
17. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: 2013 IEEE International Conference on Computer Vision. pp. 2200–2207 (2013). https://doi.org/10.1109/ICCV.2013.274
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
19. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (Nov 1995). https://doi.org/10.1145/219717.219748
20. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley mhad: A comprehensive multimodal human action database. In: IEEE WACV. IEEE (2013)
21. Perrett, T., Damen, D.: Ddlstm: dual-domain lstm for cross-dataset action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7852–7861 (2019)
22. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
23. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR. pp. 1194–1201. IEEE (2012)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Sultani, W., Saleemi, I.: Human action recognition across datasets by foreground-weighted histogram decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 764–771 (2014)
26. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
27. Yao, Y., Wang, Y., Guo, Y., Lin, J., Qin, H., Yan, J.: Cross-dataset training for class increasing object detection. arXiv preprint arXiv:2001.04621 (2020)
28. Yoshikawa, Y., Shigeto, Y., Takeuchi, A.: Metavd: A meta video dataset for enhancing human action recognition datasets. Computer Vision and Image Understanding **212**, 103276 (2021)
29. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint pattern recognition symposium. pp. 214–223. Springer (2007)
30. Zhang, J., Li, W., Ogunbona, P., Xu, D.: Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. ACM Computing Surveys (CSUR) **52**(1), 1–38 (2019)
31. Zhang, X.Y., Shi, H., Li, C., Zheng, K., Zhu, X., Duan, L.: Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In: AAAI. vol. 33, pp. 9227–9234 (2019)