Domain-adaptive Discriminative One-shot Learning of Gestures

Tomas Pfister[1], James Charles[2] and Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford
[2]Computer Vision Group, School of Computing, University of Leeds

Abstract. The objective of this paper is to recognize gestures in videos – both localizing the gesture and classifying it into one of multiple classes. We show that the performance of a gesture classifier learnt from a single (strongly supervised) training example can be boosted significantly using a 'reservoir' of weakly supervised gesture examples (and that the performance exceeds learning from the one-shot example or reservoir alone). The one-shot example and weakly supervised reservoir are from different 'domains' (different people, different videos, continuous or non-continuous gesturing, *etc.*), and we propose a domain adaptation method for human pose and hand shape that enables gesture learning methods to generalise between them. We also show the benefits of using the recently introduced Global Alignment Kernel [12], instead of the standard Dynamic Time Warping that is generally used for time alignment.
The domain adaptation and learning methods are evaluated on two large scale challenging gesture datasets: one for sign language, and the other for Italian hand gestures. In both cases performance exceeds the previous published results, including the best skeleton-classification-only entry in the 2013 ChaLearn challenge.

1 Introduction

Gesture recognition has recently received an increasing amount of attention due to the advent of Kinect and socially important applications, *e.g.* sign language to speech translation [8], becoming more tractable. However, the majority of approaches to gesture (and action) recognition rely on strongly supervised learning, which requires ground truthing large quantities of training data. This is inherently expensive and does not scale to large, evolving gesture languages with high levels of variation. As a result, several recent works have attempted to learn gestures at the other extreme – from single training examples using one-shot learning [16, 17, 19, 20, 22, 24, 33]. However, given the vast variability in how gestures are performed, and the variation in people and camera viewpoints, learning accurate, generalizable models with so little supervision is somewhat challenging, to say the least. Another avenue of work has explored learning gestures from practically infinite sources of data with weak supervision [7, 11, 23, 26, 28], *e.g.* TV broadcasts with aligned subtitles (or similarly actions from movies with aligned transcripts [1, 3, 14]). While these works have also shown promise, they

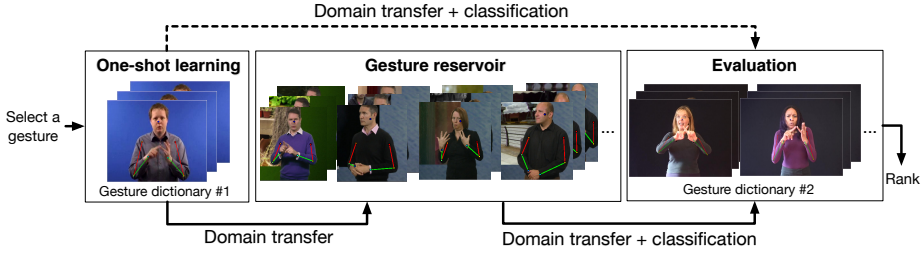


Fig. 1: **Domain-adaptive discriminative one-shot gesture learning.** Domain-adapted one-shot learning is used to obtain additional training data from a huge weakly supervised gesture repository of another domain. These new samples are used to 'boost' the one-shot learner with additional discriminative power. Evaluations are carried out under further domain adaptation on another one-shot learning dataset. Dashed line shows the **baseline** and the solid lines show the **proposed method**.

have also demonstrated the limitations of the weak supervision available for gestures today: it is so weak and noisy that it is very difficult to learn from it alone [7, 11, 28].

In this paper we show the benefit of combining these two gesture recognition approaches, one-shot learning and weakly supervised learning. The key idea is that, given suitable domain adaptations, one-shot learning can gain an enormous performance boost from utilising weakly supervised data from different domains.

Consider a common gesture recognition scenario as depicted in Fig. 1. Here we have one or more videos (one-shot learning examples; gesture 'dictionaries') showing an example of each gesture, and a huge dataset of weakly labelled videos from another domain (the 'gesture reservoir') containing some instances of the same gestures. This scenario arises naturally in gesture languages, such as sign language, which have many video dictionaries (online and on DVDs) with examples of signs, and a plentiful supply of signed data (*e.g.* sign language-translated TV broadcasts, shown in Fig. 2 (right); or linguistic research datasets). In the case of TV broadcasts the weak supervision is provided by aligned subtitles (that specify a temporal interval where the word *may* occur), though the supervision is also noisy as the subtitle word may not be signed. In the case of linguistic research datasets (and some gesture datasets [19]) the supervision is often at the video clip level, rather than a tighter temporal interval.

Our aim is to boost the performance of one-shot learning by using this large 'reservoir' of weakly supervised gestures. In the one-shot case there is strong supervision (but only one example). In the reservoir there are many examples exhibiting variations in people, expression, speed, but only weak supervision – the temporal interval is not tight. The goal is to obtain a weak classifier from one-shot learning and use it to select further examples from the reservoir. A stronger classifier can then be trained from the gesture variations and large variation of people in the reservoir. This is a form of semi-supervised learning, but here the

Fig. 2: Sample frames from the four datasets with upper body pose estimates overlaid. From left to right: BSL dictionary 1, BSL dictionary 2, ChaLearn and BSL-TV.

affinity function requires domain adaptation in going between the one-shot and gesture reservoir videos.

This is a very challenging task since the video dictionaries and gesture reservoir can be of wildly differing video domains (see Fig. 2): different size and resolutions; different people; and with gestures performed at significantly different prosody and speed. Furthermore, one domain may contain continuous gestures (*e.g.* most weakly supervised datasets) while another (*e.g.* most one-shot learning datasets) only contains gestures performed with clear breaks in-between.

In the remainder of this paper we show that not only can a weakly supervised gesture reservoir be used to significantly boost performance in gesture recognition, but also that the availability of multiple one-shot learning datasets enables evaluation of gesture recognition methods for problems where test data was previously absent. Our contributions are: (i) a method for learning gestures accurately and discriminatively from a single positive training example (in the spirit of one-shot learning and exemplar SVMs/LDAs [21, 25]); (ii) a domain adaptation method for human pose and hand shape that enables generalisation to new domains; and (iii) a learning framework for gestures that employs the recently introduced Global Alignment Kernel [12].

In the evaluations we show that our method achieves a significant performance boost on two large gesture datasets. We also release pose estimates for two gesture dictionaries[1].

[1] `http://www.robots.ox.ac.uk/~vgg/research/sign_language`

2 Domain-adaptive Discriminative One-shot Learning

In this section we first overview the learning framework, and then describe the details of the visual features (hand trajectory and hand shape) and their domain transfer which involves space and time transformations.

Figure 1 shows an overview of the learning framework. There are three domains: two gesture one-shot 'dictionaries' and one large weakly supervised 'gesture reservoir'. One dictionary is used for training, the other for testing.

The method proceeds in four steps: (1) train a discriminative one-shot gesture detector from the first dictionary, separately for each gesture; (2) that detector is then used to discover new samples of the same gesture in the weakly supervised gesture reservoir – the search for the sample is restricted to a temporal interval provided by the weak supervision; (3) these new samples are used to train what is effectively a stronger version of the original one-shot gesture classifier; and (4) this strong classifier is evaluated on a second one-shot dictionary.

2.1 Discriminative one-shot gesture learning framework

Given the two video dictionaries (one for training, the other for evaluation) and a weakly labelled gesture reservoir, let $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2$ and $\boldsymbol{\nu}$ denote their respective features (here the hand trajectories). For example, $\boldsymbol{\delta}^1 = \{\boldsymbol{\delta}^1_1, \ldots, \boldsymbol{\delta}^1_q, \ldots\}$ where $\boldsymbol{\delta}^1_q$ is a variable-length vector (depending on the length of the gesture video) of hand positions over all frames in the q^{th} gesture video of dictionary 1.

Imagine we are learning a gesture for 'snow' in BSL (shown in Fig. 3). We first train a discriminative one-shot gesture detector for 'snow' on the features of the first dictionary ($\boldsymbol{\delta}^1$). To do this, we use a time-and-space-aligned gesture kernel ψ (defined in Sect. 2.2) in a dual SVM to learn weights α from

$$\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k y_j y_k \psi(\mathbf{x}_j, \mathbf{x}_k) \quad \forall i \; 0 \leq \alpha_i \leq C \quad \sum_i \alpha_i y_i = 0 \qquad (1)$$

where we set the learning feature to $\mathbf{x} = \boldsymbol{\delta}^1$ (hand trajectories of the videos in dataset δ^1), y_i are binary video labels (1 for the 'snow' dictionary video, -1 for others), and $\psi(,)$ is the kernel. This is the one-shot learning – as an exemplar SVM [25].

In the second step, we use this model to discover new samples of 'snow' in the weakly supervised gesture reservoir (restricted to the temporal intervals provided by the weak supervision). A very large number of samples in the reservoir are scored to find gestures that are most similar to 'snow' (and dissimilar to the other gestures) in the first dictionary. This yields a vector of scores s

$$s(\boldsymbol{\nu}) = \sum_i \alpha_i y_i \psi(\mathbf{x}_i, \boldsymbol{\nu}) + b \qquad (2)$$

where $\boldsymbol{\nu}$ are the features for reservoir subsequences with a weakly supervised label 'snow'. Here, $s(\boldsymbol{\nu})$ is a vector of scores of length $|\boldsymbol{\nu}|$ (the number of samples

Fig. 3: Frames showing variation in gesture speed and prosody across two domains (top and bottom). The example gesture shown here is 'snow' in BSL, which mimics snow falling down. Although the frame rate is the same, the speed at which the gestures are produced are considerably different.

in the weakly supervised sequences of the gesture reservoir). The top scored samples represent gestures in the reservoir that are most similar to 'snow' in the the first dictionary, but with high variability in space, time and appearance (thanks to the time and space adaptations).

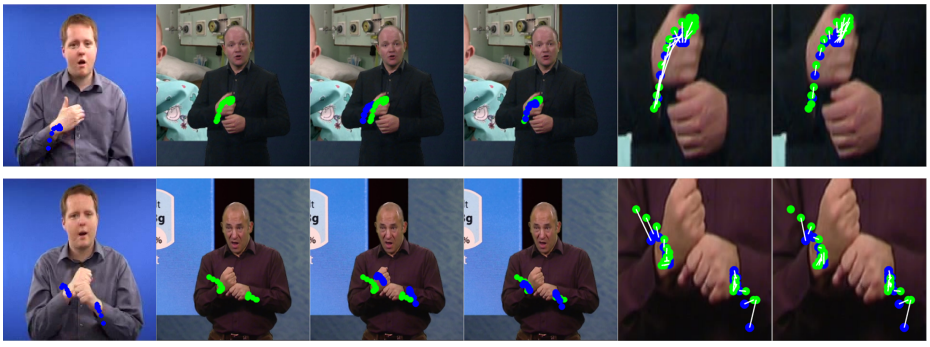
In the third step, the top samples of $s(\boldsymbol{\nu})$ (by score), along with a set of negative samples from the gesture reservoir, are used to train a stronger version of the original one-shot gesture classifier for 'snow' (training details are given in Sect. 2.4). We do this by retraining (1) with this new training set $\mathbf{x} = \boldsymbol{\nu}_{\text{retrain}}$ (of cardinality around 2,000 samples). Due to only selecting the top samples of the gesture reservoir for training, we develop resilience to noisy supervision.

In the fourth and final step, this stronger model is evaluated on the second dictionary by ranking all gesture videos using the score $s(\boldsymbol{\delta}^2)$ of the stronger classifier. This provides a measure of the strength of the classifier without requiring any expensive manual annotation.

2.2 Domain adaptations

A major challenge in gesture recognition is that not only are the gestures performed by different people with different body shapes, but the same gestures are performed at very different speeds and prosody across domains and people (see Fig. 3). We tackle this problem by measuring distance under domain adaptations in both space and time. We next discuss the domain adaptations used to define this kernel ψ.

Time alignment. Dynamic Time Warping (DTW) [30,31] is a popular method for obtaining the time alignment between two time series and measuring their similarity. However, there have been problems incorporating it into a discriminative framework (*e.g.* into kernels [2, 18, 32, 35]) due to the DTW 'distance' not satisfying the triangle inequality. As a result, it cannot be used to define a positive definite kernel. Furthermore, it is unlikely to be robust as a similarity measure as it only uses the cost of the minimum alignment.



(a) (b) (c) (d) (e) (f)

Fig. 4: **Human pose transformation across domains.** (top) BSL for 'heart' with overlaid wrist trajectory; (bottom) BSL for 'gram'. (a) wrist trajectory for domain 1, (b) trajectory for domain 2, (c) trajectory of domain 1 mapped onto domain 2 with a spatial transformation, (d) transformation with minimisation of the local position 'slack', (e) zoomed-in similarity without temporal alignment (white lines represent wrist point correspondences across the two domains), and (f) similarity with temporal alignment. As shown, the distance (proportional to the sum of the lengths of the white point correspondence lines) is significantly lower under alignment. Best seen in colour.

In this work we use a recently proposed positive definite kernel, the Global Alignment (GA) kernel [12,13]. In addition to being positive definite, it has the interesting property of considering *all* possible alignment distances instead of only the minimum (as in DTW). The kernel computes a soft-minimum of all alignment distances, generating a more robust result that reflects the costs of all paths:

$$k_{\mathrm{GA}}(\mathbf{x},\mathbf{y}) = \sum_{\pi \in \mathcal{A}(n,m)} e^{-D_{\mathbf{x},\mathbf{y}}(\pi)} \tag{3}$$

where $D_{\mathbf{x},\mathbf{y}}(\pi) = \sum_{i=1}^{|\pi|} \|x_{\pi(i)} - y_{\pi(i)}\|$ denotes the Euclidean distance between two time series $\mathbf{x},\mathbf{y}$ under alignment π, and $\mathcal{A}(n,m)$ denotes all possible alignments between two time series of length n and m. In our case $\mathbf{x},\mathbf{y}$ are two time series of spatially aligned human joint positions, *i.e.* the joint 'trajectories' of two gestures that are being compared. By incorporating all costs into the kernel we improve classification results compared to only considering the minimal cost.

Spatial alignment. Since the hands play an important role in gestures, knowing where the wrists are is valuable to any gesture recognition method. However, an issue with human joint positions is that they are not directly comparable across domains due to differences in both position, scale and human body shape. We use two simple yet effective affine transformations, one global and another local in time, that allow for translation and anisotropic scaling. This encodes a typical setup in gesture datasets, where, for a particular gesture, the persons

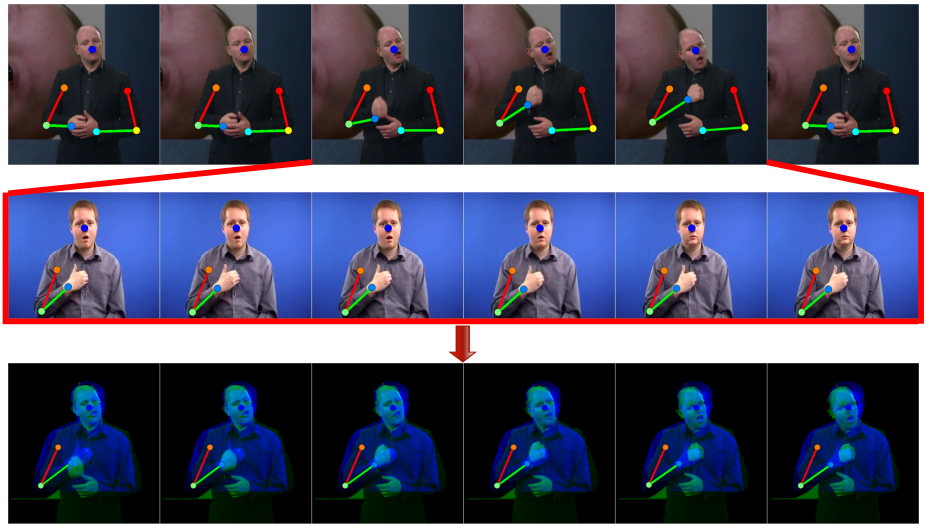


Fig. 5: **Domain adaptation in space and time. Top:** video sequence from the gesture reservoir; **Middle:** automatic time alignments to another sequence from a one-shot learning domain; **Bottom:** domain-adapted (space and time aligned) sequences, with the middle sequence overlaid on the top one. For ease of visualisation the example only uses a dominant hand, so only the dominant hand is matched (this is determined from the one-shot learning dictionary). In most cases, the transformation involves both hands.

stay at roughly the same distance from the camera (global transform), but may move slightly left or right (local transform). The global transformation learns the anisotropic scaling and translation, and the local transformation estimates an x translation, mapping into a canonical frame in which poses from different domains can be directly compared.

The global transform is computed from the median positions of the shoulders and elbows (selected since they are comparable across videos) over the whole video. The x translation is estimated locally from the median head and shoulder positions over a small temporal window (50 frames). Fig. 5 shows a visualisation of the transformation.

Even after spatial transformations, the absolute position for the gesture (relative to the torso) generally differs slightly. We solve that by adding some 'slack' to allow for slight absolute position differences. We do this by minimising the l_2 distance between wrist trajectories (of the two videos that are compared) over a small local square patch of width $u =$ (dist. between shoulders)/10. Fig. 4(c-d) shows an example of the original and corrected positions.

The composition of the global and local transformations define the spatial transformation $\boldsymbol{\phi}$, *i.e.* $\boldsymbol{\phi}(x)$ is the mapping from the trajectory in the video to the spatial canonical frame.

Final kernel. The final kernel is a composition of the the time alignment k_{GA} and spatial transformations $\boldsymbol{\phi}$, yielding the kernel $\psi(\mathbf{x}, \mathbf{y}) = k_{\mathrm{GA}}(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}))$.

2.3 Hand shape filter

As Fig. 2 demonstrates, hand shape carries much of the discriminative information in gestures, particularly in complex gesture languages such as sign language, and needs to be included in order to successfully learn gestures. We use a hand shape descriptor to discard false positives of reservoir samples where the wrist trajectories of the one-shot learning domain and the gesture reservoir match, but the hand shape is different (the similarity score is below a threshold).

Comparing hand shapes across domains is not straightforward since the domains may be of different resolution, contain different persons, lighting *etc*. Moreover, our pose estimator only provides wrist positions (not hand centres). We next describe a domain-independent, somewhat lighting-invariant hand shape descriptor that addresses these challenges.

We follow the method of Buehler *et al*. [7] where hands are first segmented, and then assigned to a cluster index. The clusters are used both to provide a distance between hand shapes and also to aid in the segmentation. To compare two hands in different domains, we assign them to their respective local domain hand cluster exemplars and measure their similarity as the distance between the HOGs of their cluster exemplars (shown in Fig. 6).

In detail, GraphCut [5,29] is used for an initial segmentation (with skin colour posteriors obtained from a face detector), and the segmented hands are represented using HOG features (of dimensionality $15 \times 15 \times 31$). The segmentation is performed within a box defined by an estimate of hand centre position (based on the elbow-wrist vector). Hand exemplars are then formed by clustering HOG vectors for examples that are *far away* from the face using k-means ($K = 1000$). These are effectively 'clean' hand clusters, without face regions in the foreground segmentation. For an input image, HOG vectors are matched to their nearest hand cluster, resulting in a 'cleaned' segmentation of the hand.

2.4 Implementation details

Learning framework. For each word, the positive training samples are obtained from the top ranked positive samples of each reservoir video. If there are w_c occurrences of the word in the subtitles of a reservoir video, then the top $5w_c$ positive samples are used – note, no non-maximum suppression is used when sliding the classifier window so there are multiple responses for each occurrence. The number of positives is capped at 1,000, and 1,000 randomly sampled reservoir gestures are used as negatives.

Time alignment. We use the dual formulation of SVMs since our space and time alignment method provides the alignments as kernels, not in feature space (so primal optimisation is not suitable).

Hands. We precompute a $K \times K$ hand distance matrix offline for any pair of one-shot learning domain and gesture reservoir videos. At runtime, the comparison

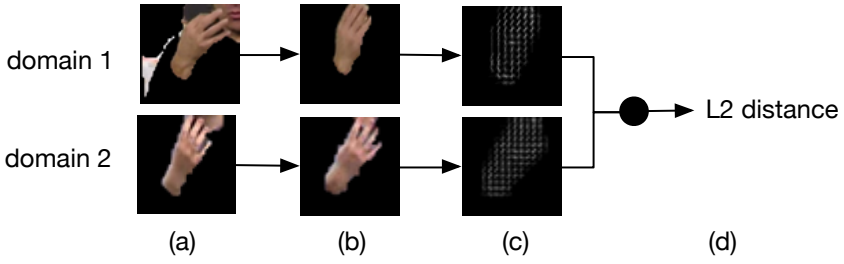


Fig. 6: **Hand shape descriptor.** (a) Badly segmented hands (due to overlap with skin) in two domains, (b) hands assigned to their hand cluster exemplars, (c) HOG of size-normalised exemplars, and (d) hands are compared across domains in terms of l_2 distance between the HOGs of the hand exemplars.

of two gestures is reduced to looking up the distance in the matrix for each pair of time-aligned frames, and summing up the distances.

Computation time. The computation times for the preprocessing steps are: pose estimation 0.4s/frame; hand segmentation 0.1s/frame. Time alignment is approx 0.001s per gesture pair, or 1,000s for a 1000×1000 kernel matrix. Other costs (*e.g.* spatial alignments, SVM training and testing, subtitle preprocessing *etc.*) are negligible in comparison (a few seconds per gesture/video).

3 Datasets

Four datasets are employed in this work: a sign language dataset extracted from TV broadcasts; two sign language dictionaries; and a dataset of Italian hand gestures. Samples from each gesture dataset are shown in Figure 2.

3.1 BSL-TV sign language dataset

This contains 155 hours of continuous British Sign Language (BSL), performed by 45 signers, with over 1,000 different continuous signs per 1hr video (and an estimated over 4,000 different signs in total). This dataset is particularly challenging to use as the supervision (in the form of subtitles) is both weak and noisy. It is weak as the subtitles are not temporally aligned with the signs – a sign (typically 8–15 frames long) could be anywhere in the overlapping subtitle video sequences (typically 400 frames). Furthermore, it is noisy as the occurrence of a word in the subtitle does not always imply that the word is signed (typically the word is signed only in 20–60% of the subtitle sequences). Furthermore, the gestures are continuous (no breaks between signs) and contain considerable variation (in terms of gesturing speed, signers, and regional gesturing differences).

Data preprocessing. Given a word, the subtitles define a set of subtitle sequences in which the word occurs (8–40 sequences depending on how many times the word occurs), each around 15s long. As in [28], we slide a window along each

subtitle sequence (fixed to 13 frames since that captures the majority of the gestures; gestures shorter than 13 frames are 'cropped' by the time alignment). This produces in total roughly 400 temporal windows per subtitle sequence, which are reduced to 100 candidate temporal windows per subtitle sequence using the method of [28], where only windows in which the signer also mouths the sign are considered. Upper body joint tracks are obtained automatically using the Random Forest regressor of Charles *et al.* [9, 10, 27].

3.2 Two BSL dictionary datasets

The video dictionaries are: 'Signstation' (BSL dictionary 1) [6] and 'Standard BSL dictionary' (BSL dictionary 2) [4]. The first contains 3,970 videos (total 2.5 hours), one for each word; and the second contains 3,409 videos (total 3 hours), and covers 1,771 words (the majority of words signed in one or more regional variation). BSL dictionary 1 contains a single signer, whereas BSL dictionary 2 contains multiple signers and multiple regional variations. There is no overlap of signers between the two dictionaries. The two datasets contain different sets of gestures, and intersect (*i.e.* have common words) only for a subset of these.

Data preprocessing. Upper body joint tracks are obtained automatically using the method of Charles *et al.* [9]. In order to effectively use this data (as one-shot training and testing material), it is first necessary to find the pairs of gestures (across the dictionaries) that are the same. This is made difficult by the fact that the dictionaries contain different regional variations of the same gestures (*i.e.*, we cannot simply assume gestures with the same English word label are the same). We therefore need to look for visual similarity as well as the same English word label. We automatically find a subset of words pairs of the same gesture performed the same way by computing a time-and-space aligned distance (see Sect. 2.2) from upper body joint positions for all gesture pairs of the same word, selecting pairs with distance below a threshold (set from a small manually labelled set of pairs). This list of pairs is manually verified and any false matches (mainly due to incorrect pose estimates) are filtered away. This results in 500 signs in common between the two dictionaries.

3.3 ChaLearn gesture dataset

The fourth dataset is the ChaLearn 2013 Multi-modal gesture dataset [15], which contains 23 hours of Kinect data of 27 persons performing 20 Italian gestures. The data includes RGB, depth, foreground segmentations and Kinect skeletons. The data is split into train, validation and test sets, with in total 955 videos each lasting 1–2min and containing 8–20 non-continuous gestures. In comparison, each 15s subtitle sequence in BSL-TV contains 30–40 gestures (in which the gesture may or may not occur), and are continuous, so they cannot be easily segmented.

4 Evaluation

Experiments are conducted on the four datasets introduced in Sect. 3. See our website for example videos showing qualitative results.

4.1 One-shot detection of gestures in the gesture reservoir

Here we evaluate the first main component of our method, *i.e.* how well can we spot gestures in the gesture reservoir given a one-shot learning example? We compare it to previous work [28] that took a different approach (based on Multiple Instance Learning, MIL) to extracting gestures from weakly supervised gesture datasets. We show that we vastly outperform previous work on the same data with our conceptually much simpler one-shot learning approach.

Manual ground truth. The test dataset, a six hour subset of BSL-TV, is annotated for six gestures (bear, gram, heart, reindeer, snow and winter), with on average 18 occurrences for each gesture, and frame-level manual ground truth from Pfister *et al.* [28] (where we spent a week to label 41 words frame-by-frame). A benefit of the domain adaptation method is that it renders this expensive manual labelling unnecessary, since the training and test sets no longer need to be of the same domain. This enables the use of supervised datasets from other domains for testing (as done in the next experiment with a dictionary).

Task. The task for each of the six gestures is, given one of the 15s temporal windows of continuous gestures, to find which windows contain the target gesture and provide a ranked list of best estimates. Only about 0.5s out of 15s actually contain an instance of the gesture; the remainder contain other gestures. A gesture is deemed 'correct' if it overlaps at least 50% with ground truth.

Results. Precision-recall curves for the gestures are given in Fig. 8(left). As shown, thanks to our domain-adapted one-shot learning method, we vastly outperform the approach of Pfister *et al.* which uses MIL to pick temporal windows that occur frequently where subtitles say they should occur (and infrequently elsewhere). In contrast, [28] do not use any direct supervision (which our domain-adapted one-shot learner provides). This shows very clearly the high value of one-shot learning for extracting additional gesture training data. In fact, our method can complement Pfister *et al.*'s by using it for gestures that exist in our one-shot learning domains, and using [28] for other gestures.

4.2 Domain-adapted discriminative one-shot gesture learner

In this key experiment we evaluate our discriminative one-shot learning method trained on the 155 hour BSL-TV gesture reservoir. The method is evaluated on a second one-shot learning domain ('BSL dictionary 2') on the same gestures as in the first one-shot learning domain ('BSL dictionary 1'), but in a different domain, signed at different speeds by different people. The second dictionary is used as the testing set to reduce annotation effort (the BSL-TV reservoir does not come with frame-level labels). Sect. 3 explains how these cross-dictionary gesture 'pairs' that contain the same sign signed the same way are found.

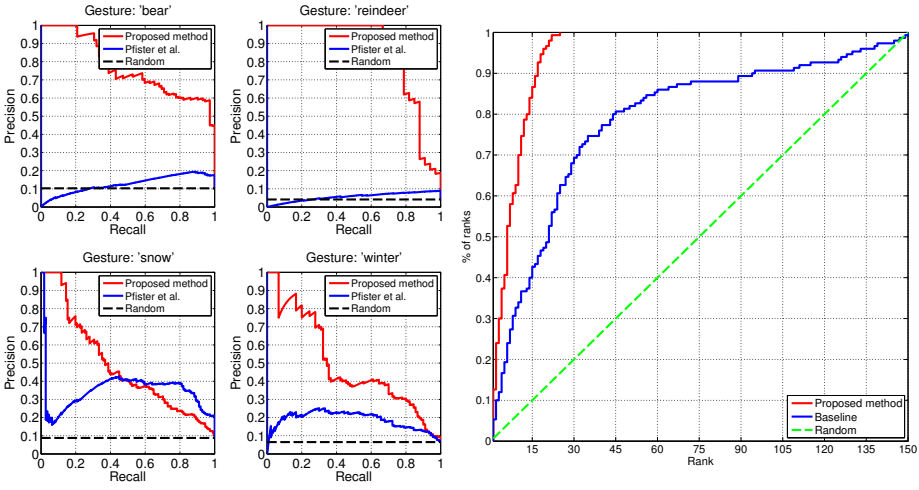


Fig. 8: **Left:** Gesture spotting accuracy on the gesture reservoir for BSL-TV, with a comparison to Pfister *et al*. [28]. PR curves are for four gestures with ground truth (see website for curves for remaining gestures). **Right:** Gesture classifier accuracy evaluated on gesture dictionary 'BSL dictionary 2'. The graph shows the cumulative distribution function of the ranks for the baseline and our proposed method. For example, 87% of gestures are ranked within the top 15.

Baseline one-shot learner. We compare our method to an enhanced one-shot learning method trained on one one-shot learning domain ('BSL dictionary 1') and tested on the other ('BSL dictionary 2'), without any weakly supervised additional training data from the gesture reservoir (as shown at the top of Fig. 1). The method uses the time and space domain adaptations.

Training and testing set. The cross-dictionary gesture 'pairs' that contain the same sign signed the same way (found as explained in Sect. 3) define an 'in-common' set of 500 signs. The training set consists of the 150 gestures from BSL dictionary 1 from the in-common set for which a sufficient number of examples exist in the BSL-TV gesture reservoir (set to at least 16 subtitle occurrences). The testing set consists of the same set of 150 gestures from BSL dictionary 2.

Test task & evaluation measure. Each of the 150 training gestures is evaluated independently. For each gesture, the gesture classifier is applied to all 150 test gestures, one of which contains the correct gesture. The output of this step is, for each gesture classifier, a ranked list of 150 gestures (with scores). The task is to get the correct gesture first. Each gesture classifier is assigned the rank of the position in the 150-length list in which the correct gesture appears.

Results. Fig. 8(right) shows a cumulative distribution function of the ranks for the baseline and our proposed method using the gesture reservoir. We clearly see that, although the baseline ranks 66% of the gestures within the first top 60, learning from the reservoir beats it, with all gestures ranked within the first 25,

13% as rank 1, 41% within the first 5, and 70% within the first 10. We believe this is due to the high training data variability that the additional supervision from the gesture reservoir provides (from multiple persons, with gestures performed with many different speeds *etc.*).

There are two principal failure modes: first, the majority of gestures with ranks above 15 are due to several gestures out of the test gestures having very similar hand trajectories and hand shapes. With an already challenging discrimination problem, this causes confusions when the gesture in the evaluation set is performed very differently from any gesture in the training reservoir. The other major problem source is inaccurate pose estimates, which results in inaccurate hand trajectory and hand shape estimates.

Component evaluation. Each of the components of our method is evaluated by switching one off at a time, and reporting rank-15 accuracy. Changing the time alignment method from global alignment to DTW decreases the rank-15 accuracy from 87% to 51%; switching off hand shape lowers it to 72%; and switching off time alignment for the one-shot dictionary learner drops it to 46%.

Our method works despite the domain adaptations between the one-shot dictionaries and weakly supervised datasets being very challenging: different resolutions, settings, people, gesture speed and regional variations; and one domain (the one-shot dictionaries) containing non co-articulated gestures (*i.e.* having breaks between gestures) whereas others (the gesture reservoir) only contain continuous gestures. To add to all of this, the supervision in the weakly supervised datasets is very weak and noisy. Despite all these challenges, we show a considerable performance boost. We consistently outperform the one-shot learning method, and achieve much higher precision and recall than previous methods in selecting similar gestures from the gesture reservoir using weak supervision.

4.3 Comparison on ChaLearn multi-modal dataset

On the ChaLearn dataset we define the one-shot learning domain as the training data for one person, and keep the remaining training data (of the 26 other persons) as the unlabelled 'gesture reservoir'. Only Kinect skeletons are kept for the reservoir. We compare this setup to using all the ground truth for training.

Task. The task here is, given a test video (also containing distractors), to spot gestures and label them into one out of 20 gesture categories.

Audio for gesture segmentation. Gestures only appear in a small subset of the dataset frames, so it makes sense to spot candidate windows first. To this end we use the same method as the top entries in the ChaLearn competition: segment gestures using voice activity detection (the persons pronounce the word they gesture). However, we do not use audio for classification since our purpose is to evaluate our vision-based classifier. We therefore compare only to methods that do not use audio for classification but only use it for segmentation (including the winner's method without audio classification).

Baseline, our method & upper bound. The baseline is domain-adapted one-shot learning (where training data comes from a single person from the 27 person training dataset; we report an average and standard deviation over each

possible choice). This is compared to our method that uses the one-shot learner to extract additional training data from the unlabelled 'gesture reservoir'. The upper bound method uses all training data with manual ground truth.

Experiment overview. In Experiment 1 we compare in detail to the competition winner [34] with the same segmentation method (audio), using only skeleton features for classification, and evaluating in terms of precision and recall on the validation set. In Experiment 2 we compare to competition entrants using the standard competition evaluation measure on test data, the Levenshtein distance $L(R,T)$, where R and T are ordered lists (predicted and ground truth) corresponding to the indices of the recognized gestures (1–20); distances are summed over all test videos and divided by the total number of gestures in ground truth.

Results for Experiment 1. Our method achieves very respectable performance using a fraction of the manually labelled data that the other competition entrants use. The competition winner's method gets Precision $P = 0.5991$ and Recall $R = 0.5929$ (higher is better) using skeleton features for classification [34]. Using this exact same setup and test data, our baseline one-shot learner achieves $P = 0.4012$ (std 0.015) and $R = 0.4162$ (std 0.011) – notably by only using a single training example, whereas the winner used the whole training set containing more than 400 training examples per class. Our results are improved further to $P = 0.5835$ (std 0.021), $R = 0.5754$ (std 0.015) by using gestures extracted from the gesture reservoir, still only using one manually labelled training example per gesture. Using the whole training set yields $P = 0.6124$, $R = 0.6237$.

Results for Experiment 2. In terms of Levenshtein distance, our method improves from the baseline 0.5138 (std 0.012) to 0.3762 (std 0.015) (lower is better). With only a single training example (two orders of magnitude less manually labelled training data than other competition entries) we achieve similar performance to the best method using skeleton for classification ('SUMO', score 0.3165 [15]), and using the full training set we outperform them at 0.3015.

5 Conclusion

We have presented a method that utilises weakly supervised training data containing multiple instances of a gesture to significantly improve the performance of a gesture classifier. Another benefit of our framework with two dictionary datasets is that it lets us avoid a very expensive laborious task that has been a big issue for weakly supervised gesture recognition: large-scale evaluation. Our approach is applicable to gesture recognition in general – where the upper body and hands are mostly visible, and the person is communicating with gestures.

Acknowledgements: We are grateful to Patrick Buehler and Sophia Pfister for help and discussions. Financial support was provided by EPSRC grant EP/I012001/1.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE PAMI 32(2), 288–303 (2010)
2. Baisero, A., Pokorny, F.T., Kragic, D., Ek, C.: The path kernel. In: ICPRAM (2013)
3. Bojanowski, P., Bach, F., , Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proc. ICCV (2013)
4. Books, M.: The standard dictionary of the British sign language. DVD (2005)
5. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: Proc. ICCV (2001)
6. Bristol Centre for Deaf Studies: Signstation. `http://www.signstation.org`, [Online; accessed 1-March-2014]
7. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: Proc. CVPR (2009)
8. Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, M.: Sign language recognition and translation with Kinect. In: Proc. Int. Conf. Autom. Face and Gesture Recog. (2013)
9. Charles, J., Pfister, T., Everingham, M., Zisserman, A.: Automatic and efficient human pose estimation for sign language videos. IJCV (2013)
10. Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Domain adaptation for upper body pose tracking in signed TV broadcasts. In: Proc. BMVC (2013)
11. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Proc. CVPR (2009)
12. Cuturi, M.: Fast global alignment kernels. In: ICML (2011)
13. Cuturi, M., Vert, J., Birkenes, Ø., Matsui, T.: A kernel for time series based on global alignments. In: ICASSP (2007)
14. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: Proc. CVPR (2009)
15. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Sigal, L., Argyros, A., Sminchisescu, C.: Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In: ACM MM (2013)
16. Fanello, S., Gori, I., Metta, G., Odone, F.: Keep it simple and sparse: real-time action recognition. J. Machine Learning Research 14(1), 2617–2640 (2013)
17. Farhadi, A., Forsyth, D., White, R.: Transfer learning in sign language. In: Proc. CVPR (2007)
18. Gaidon, A., Harchaoui, Z., Schmid, C.: A time series kernel for action recognition. In: Proc. BMVC (2011)
19. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H., Hamner, B.: Results and analysis of the ChaLearn gesture challenge 2012. In: Proc. ICPR (2013)
20. Guyon, I., Athitsos, V., Jangyodsuk, P., Hamner, B., Escalante, H.: ChaLearn gesture challenge: Design and first results. In: CVPR workshops (2012)
21. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: Proc. ECCV (2012)
22. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: Proc. ICCV (2007)
23. Kelly, D., McDonald, J., Markham, C.: Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. Trans. Systems, Man, and Cybernetics 41(2), 526–541 (2011)

24. Krishnan, R., Sarkar, S.: Similarity measure between two gestures using triplets. In: CVPR Workshops (2013)
25. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: Proc. ICCV (2011)
26. Nayak, S., Duncan, K., Sarkar, S., Loeding, B.: Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. J. Machine Learning Research 13(1), 2589–2615 (2012)
27. Pfister, T., Charles, J., Everingham, M., Zisserman, A.: Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In: Proc. BMVC (2012)
28. Pfister, T., Charles, J., Zisserman, A.: Large-scale learning of sign language by watching TV (using co-occurrences). In: Proc. BMVC (2013)
29. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. In: Proc. ACM SIGGRAPH (2004)
30. Sakoe, H.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing (1978)
31. Sakoe, H., Chiba, S.: A similarity evaluation of speech patterns by dynamic programming. In: Nat. Meeting of Institute of Electronic Communications Engineers of Japan (1970)
32. Shimodaira, H., Noma, K., Nakai, M., Sagayama, S.: Dynamic time-alignment kernel in support vector machine. In: NIPS (2001)
33. Wan, J., Ruan, Q., Li, W., Deng, S.: One-shot learning gesture recognition from RGB-D data using bag of features. J. Machine Learning Research 14(1), 2549–2582 (2013)
34. Wu, J., Cheng, J., Zhao, C., Lu, H.: Fusing multi-modal features for gesture recognition. In: ICMI (2013)
35. Zhou, F., De la Torre, F.: Generalized time warping for multi-modal alignment of human motion. In: Proc. CVPR (2012)