Domain-adaptive Discriminative One-shot Learning of Gestures

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

[1] University of Oxford

[2] University of Leeds

Objective and Contributions

Recognise gestures in videos – both localising the gesture and classifying it into one of multiple classes.

- **Learning gestures from one-shot+weak supervision**
- **Domain adaptation for human pose and hand shape**
- **Benefits of using Global Alignment kernels**

Motivation

Most gesture recognition methods rely on **strong supervision**

⚠ Manual annotation is expensive & does not scale

***Alternative 1:* One-shot supervision** (a single training example)

⚠ Generalisation very challenging

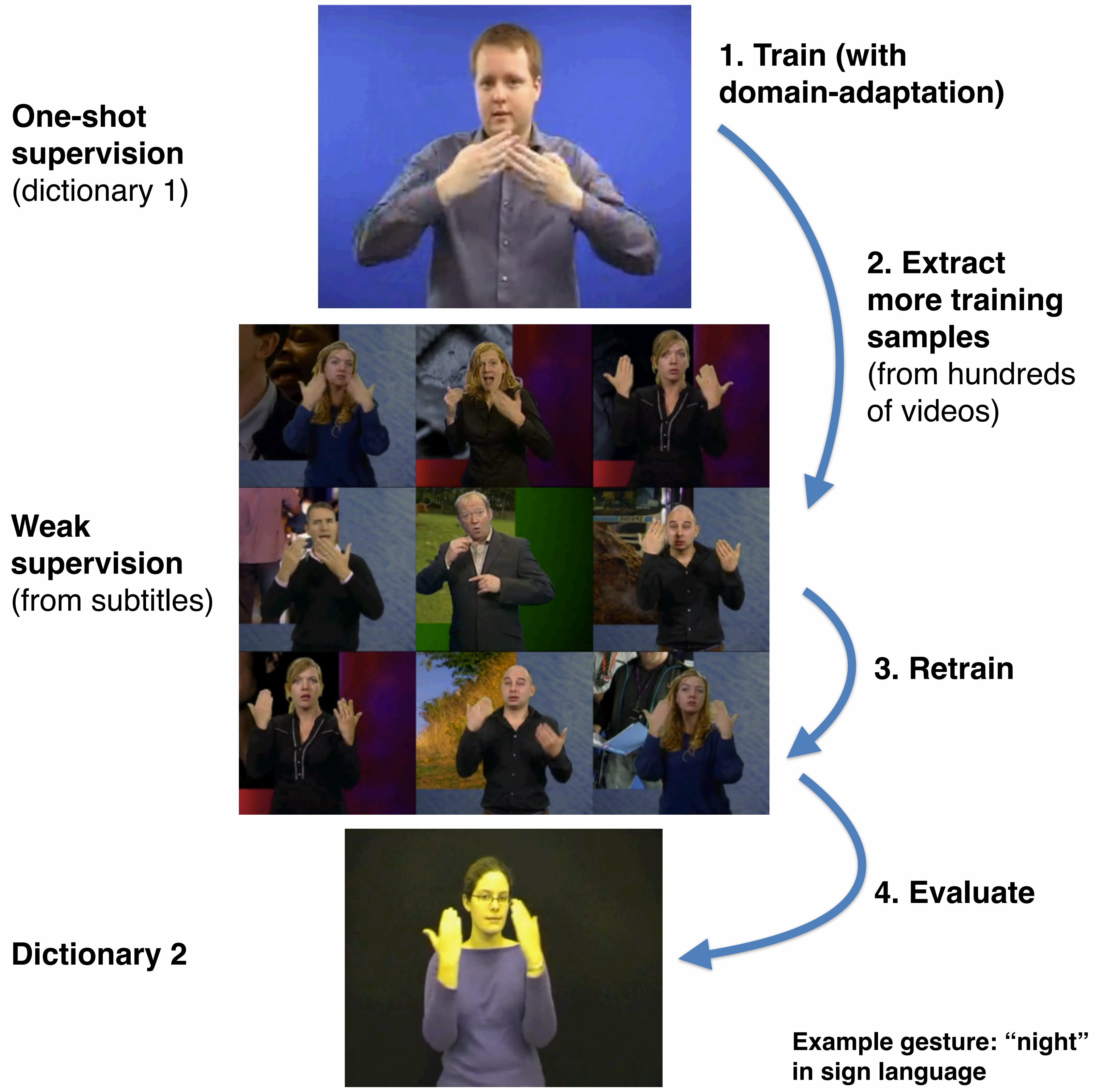
***Alternative 2:* Weak supervision** (e.g. subtitles of TV broadcasts)

⚠ Often too weak and noisy to learn good models

***Our work:* Combine one-shot + weak supervision**

😎 No annotation needed & Generalising models

Overview



One-shot supervision (dictionary 1)

1. Train (with domain-adaptation)

2. Extract more training samples (from hundreds of videos)

Weak supervision (from subtitles)

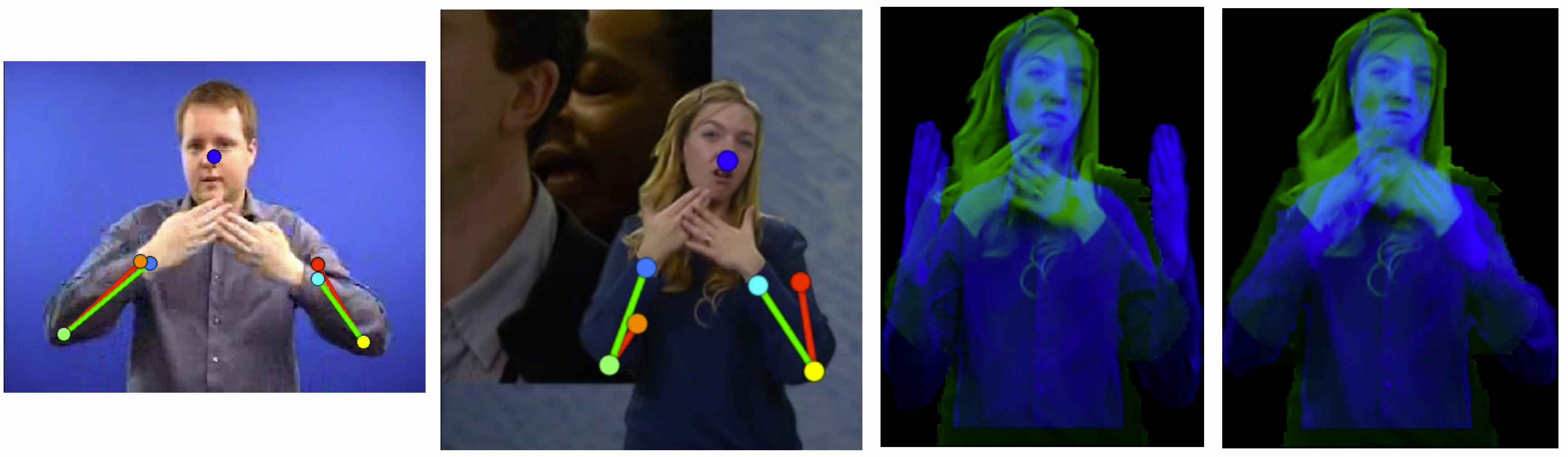
3. Retrain

4. Evaluate

Dictionary 2

Example gesture: "night" in sign language

Step 1: Domain transfer



Strongly supervised domain — Weakly supervised domain — Space-aligned — Space and time-aligned

Space alignment:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

Time alignment [1]:

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

Soft-min of **all** alignment distances (vs min in DTW) → More robust

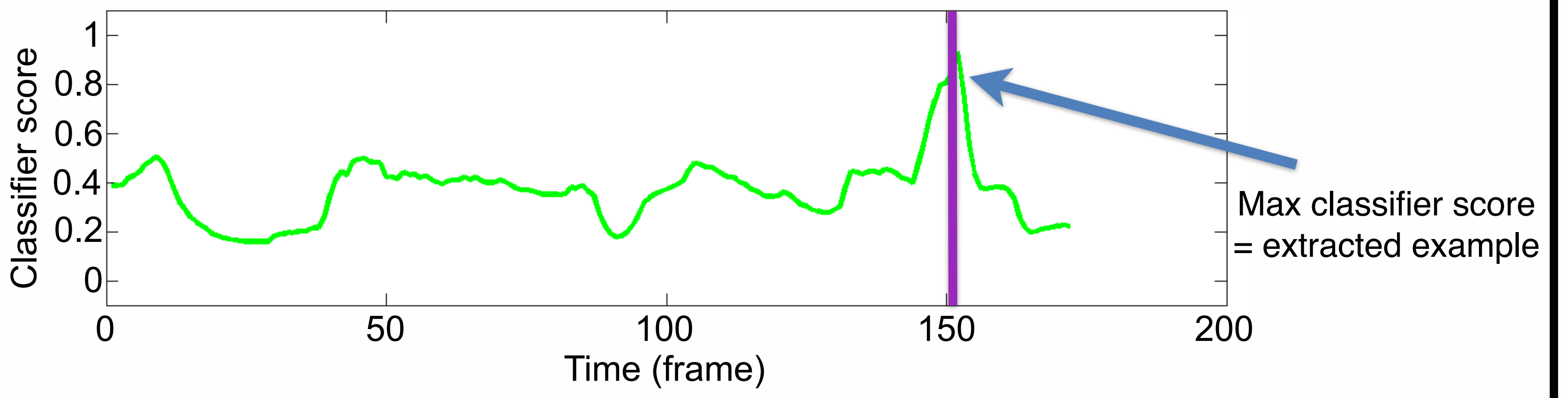
[1] M. Cuturi, Fast Global Alignment Kernels, ICML 2011

Step 2: Use one-shot supervision to find more examples

One-shot supervision for "night"

seeing something as dangerous as this in the forest at night time.

Weak supervision for "night"



Max classifier score = extracted example

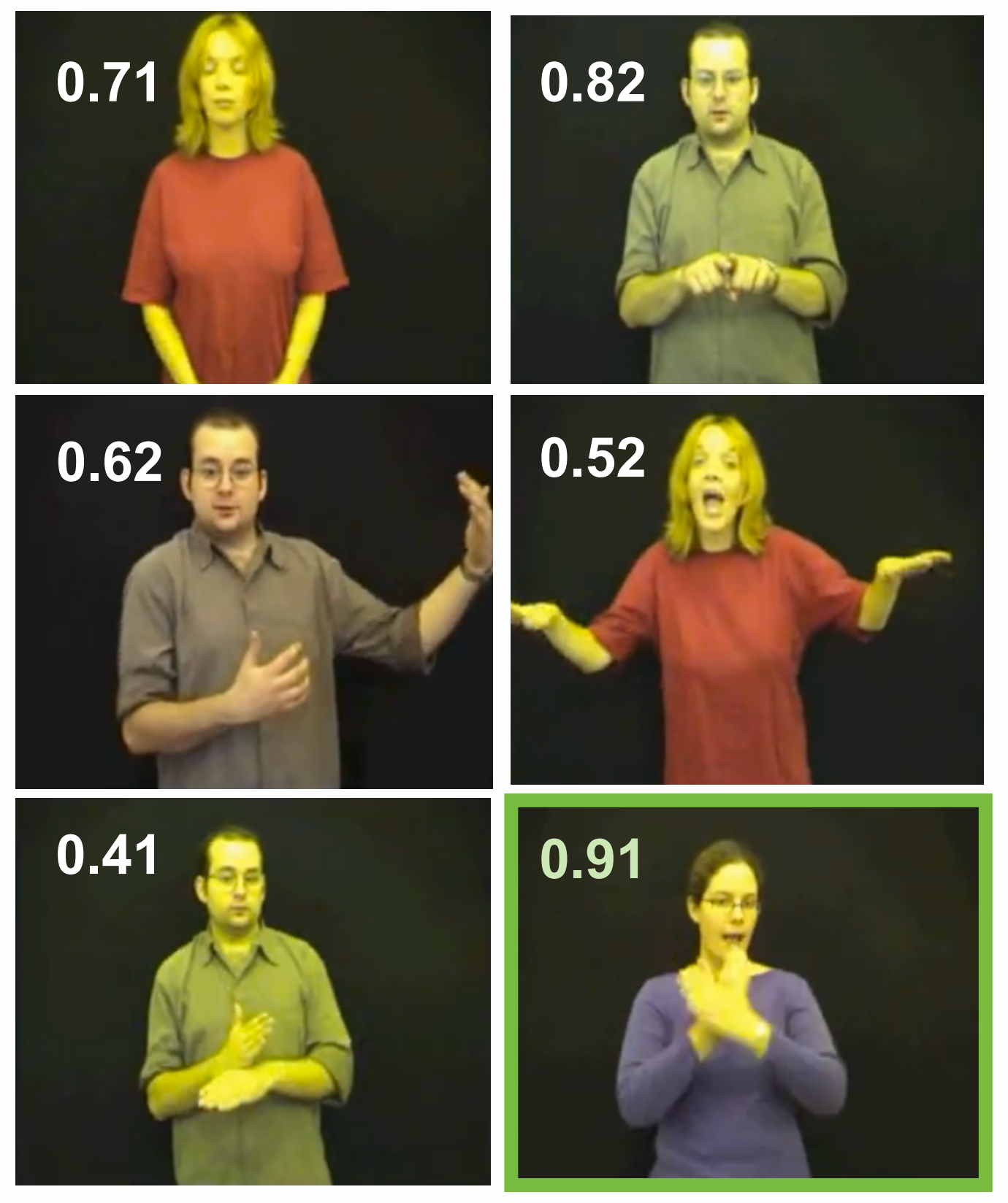
Sliding temporal window classifier on weakly supervised video

Output for "night" (from multiple weakly supervised videos)

Steps 3+4: Retrain+evaluate classifier on new examples

Extracted examples from weakly supervised domain

Global alignment SVM →

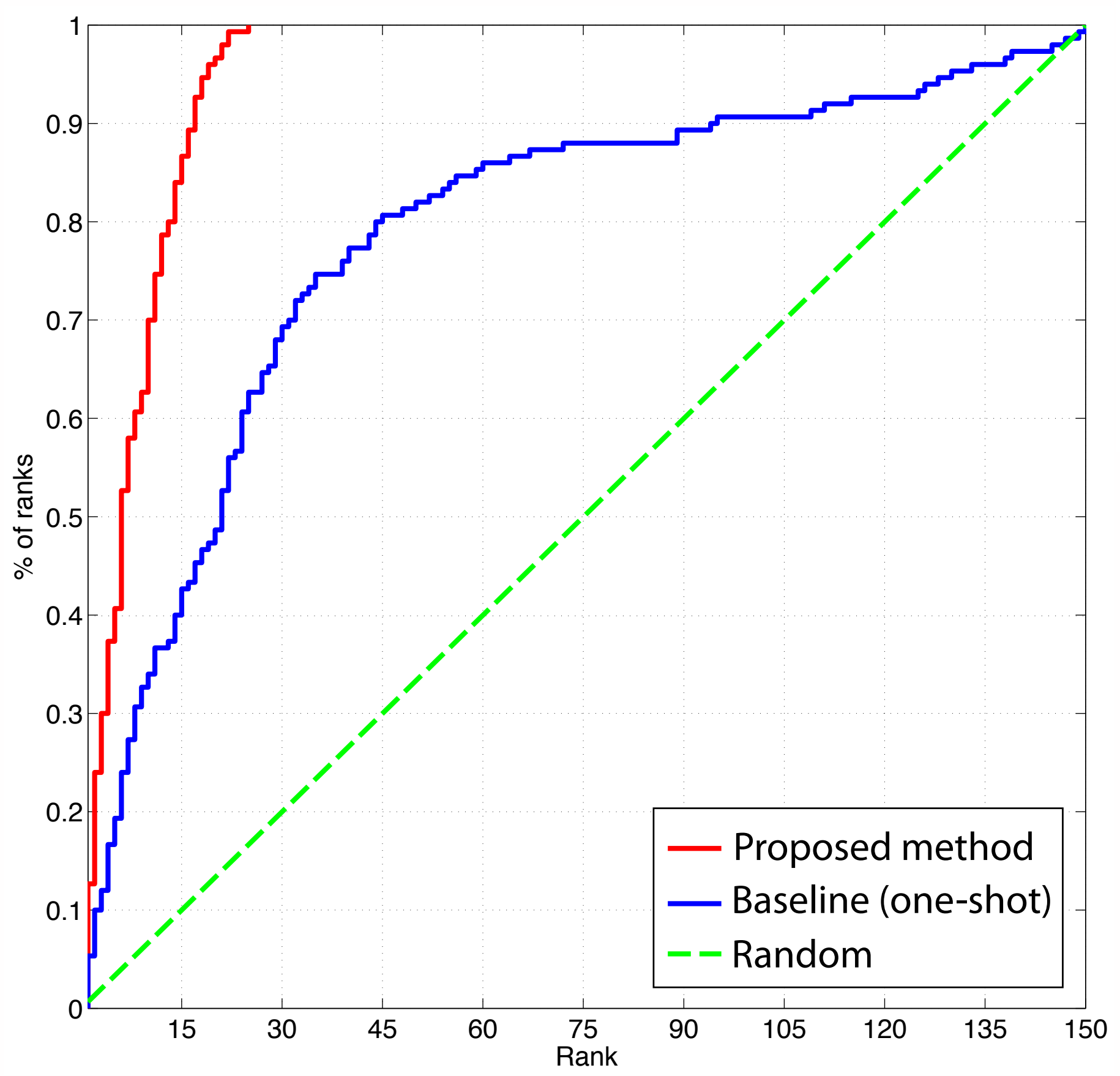


Test set

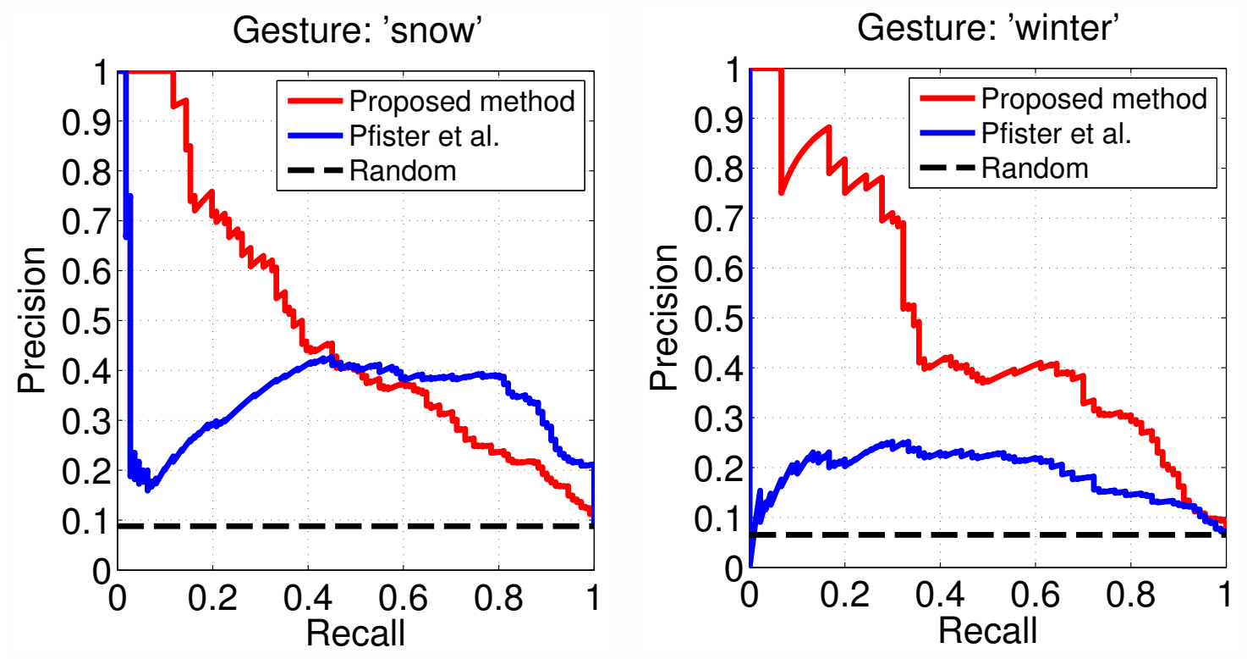
Experiments

BSL sign language dataset (155 hrs of video!)

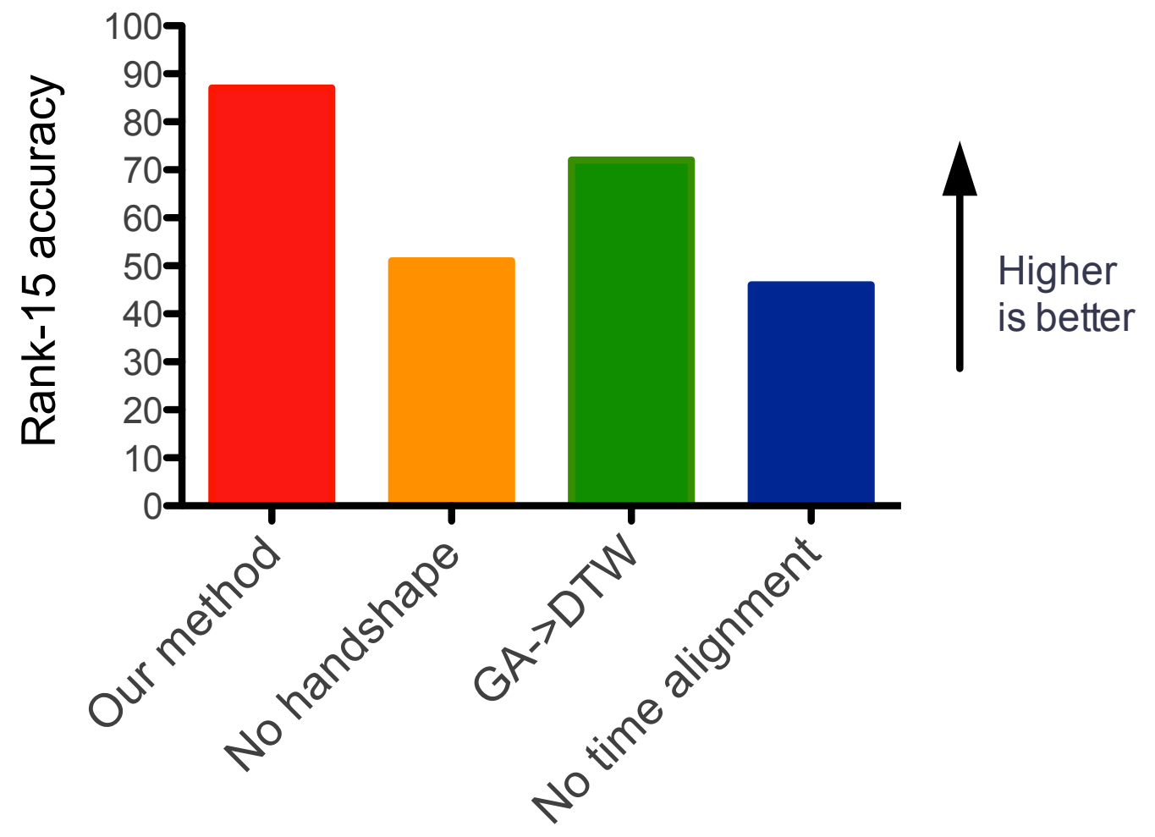
150 automatically learnt signs



Comparison to previous work

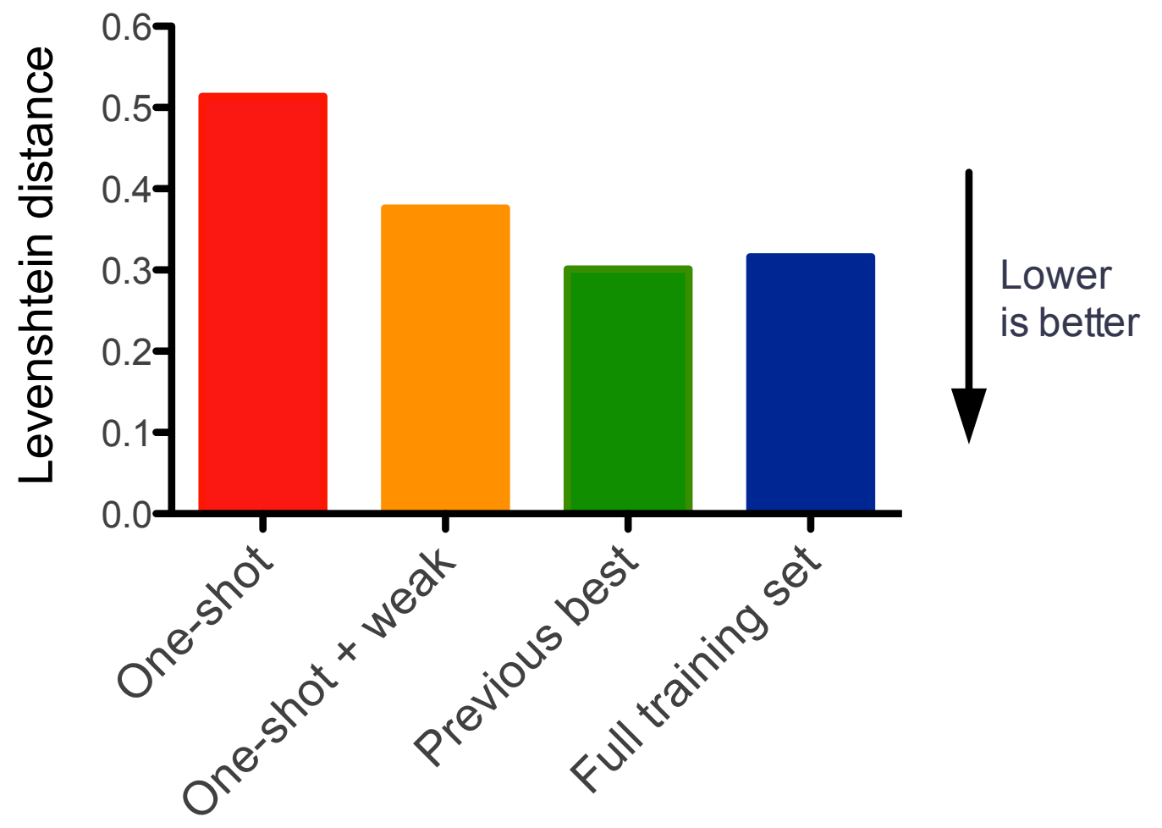


Component evaluation



Chalearn 2013 dataset

Comparison to previous work


Levenshtein distance
One-shot
One-shot + weak
Previous best
Full training set
Lower is better