

Action Recognition from Weak Alignment of Body Parts

Minh Hoai¹

<http://www.robots.ox.ac.uk/~minhhoai/>

L'ubor Ladický²

<http://www.inf.ethz.ch/personal/ladicky/>

Andrew Zisserman¹

<http://www.robots.ox.ac.uk/~az/>

¹ Visual Geometry Group

Department of Engineering Science

University of Oxford,

Oxford, UK

² ETH Zürich

Zürich, Switzerland

The objective of this paper is to recognize human actions in still images. The contribution of this work is a novel framework for obtaining weak alignment of human body-parts to improve the recognition performance. Our framework implicitly exploits physical constraints of human body parts (e.g., heads are above necks, hands are attached to forearms). It uses the locations of some detected body parts to aid the alignment of some others. Specifically, we demonstrate the benefit of our framework for computing registered feature descriptors from automatically detected upper bodies and silhouettes. Fig. 1 illustrates the benefits of our approach over the grid-alignment approach.

Given the bounding box of a human, we approximate the human body by a set of deformable rectangular parts, which is similar to a DPM [1]. The goal is to align these rectangular parts between two images, referred to as *reference* and *probe* images. We formulate the problem as a minimization of a deformation energy between the parts of the reference (which are fixed as a default grid formation) and those of the probe (which deform to best match those of the reference). The energy encourages the parts to overlap the silhouette and upper body in a consistent way (between reference and probe) whilst penalizing severe deformations. The energy is defined for a configuration of parts, and it is formulated as the sum of unary and pairwise terms. Consider aligning a human specified by a bounding box \mathbf{b} in the probe image \mathbf{I} to another human specified by the bounding \mathbf{b}^{ref} in the reference image \mathbf{I}^{ref} . Let $\mathbf{p}_1^{ref}, \dots, \mathbf{p}_k^{ref}$ be the default configuration of parts for the reference image at the bounding box \mathbf{b}^{ref} . We consider the following energy function for a configuration of parts $\mathbf{p}_1, \dots, \mathbf{p}_k$ of a probe image \mathbf{I} :

$$E(\{\mathbf{p}_i\}) = \sum_{i=1}^k \|\phi(\mathbf{I}, \mathbf{p}_i) - \phi(\mathbf{I}^{ref}, \mathbf{p}_i^{ref})\|^2 + \lambda \sum_{i=1}^k \|\psi(\mathbf{p}_i, par(\mathbf{p}_i)) - \psi_i^{def}\|^2. \quad (1)$$

The above energy function factors into a sum of local and pairwise energies. $\phi(\mathbf{I}, \mathbf{p}_i)$ is the feature vector computed at the location specified by part \mathbf{p}_i of image \mathbf{I} . In this work, it is a vector of two components. The first component is the proportion of pixels inside \mathbf{p}_i that belong to the detected upper body, and the second component is the proportion of pixels inside \mathbf{p}_i that belong to the human segmentation. $par(\mathbf{p}_i)$ is the parent of \mathbf{p}_i ; the parent of the root part is the provided bounding box \mathbf{b} . ψ is the function that computes the relative displacement of a part and its parent. ψ_i^{def} is the displacement computed for the default configuration of parts. The energy for a configuration of parts is given by the difference of each part at its respective location w.r.t. the corresponding part in the reference image (data term) plus a deformation cost that depends on the relative positions of each part w.r.t. the parent (spatial prior).

We align an image with a set of training (or reference) images as follows. We first divide the training images into three roughly equal subsets, based on the aspect ratios of the provided person bounding boxes. Given a probe image (either training or testing), we determine the subset that has similar aspect ratio, and compute the matching energy between the probe image and every training image in the subset. The matching energy is the difference (in the occupancy of silhouette and upper body) between the two default configurations of parts, as defined in Eq. 1. The m training images that yield the lowest matching energies, referred to as m nearest neighbors, are used as the references for aligning the probe image. This produces m configurations of parts for the probe image, defining its deformation space.

The alignment of a probe image w.r.t. its nearest neighbors can be used to compute an improved feature descriptor for any type of feature, including HOG and color. For example, consider a feature descriptor in which a HOG template is computed for each part. Using our approach, for each of the nearest neighbors, the HOG template can be computed at the

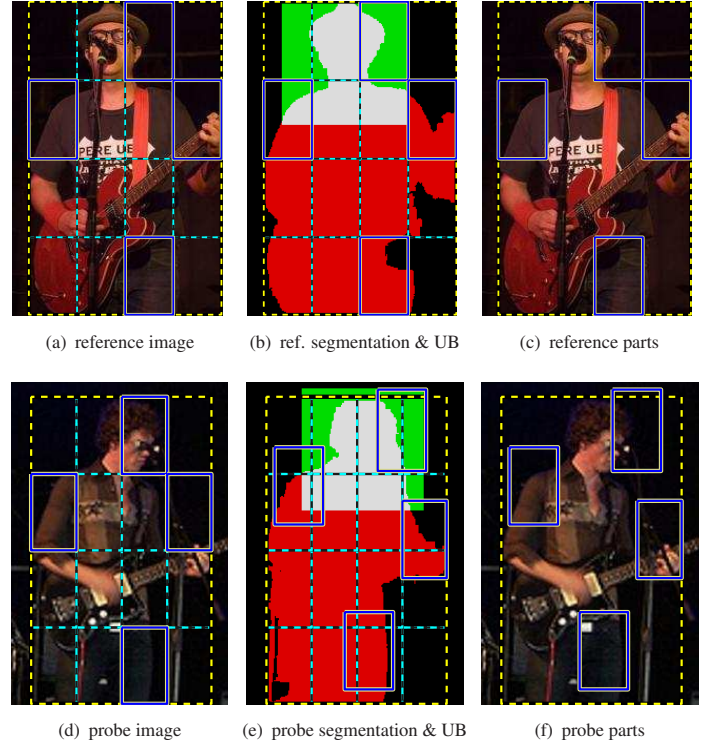


Figure 1: **Aligning body parts for action recognition.** (a) & (d): the alignment induced by a regular grid is not suited for registering body parts (c.f., solid blue boxes). The geometric constraints provided by the silhouettes and upper bodies ((b) & (e)) lead to a good alignment of parts. (c) & (f): alignment results – the translated parts are better aligned with the reference parts (e.g., the rightmost blue boxes both correspond to a hand holding the guitar fretboard).

deformed configuration of parts. We pool the HOGs for each corresponding part by averaging. The process can be thought as alignment-informed jittering.

Human silhouettes are obtained using a foreground/background segmentation algorithm. This algorithm is based on a joint energy minimization framework [2] that consists of energy potentials from a pose model, a color model, and texture classifiers. To localize the upper body, the Calvin upper-body detector is used.

We train a kernel SVM for each action class. The SVM kernel is a convex combination of base kernels, which capture different visual cues: HOG, SIFT, color, pose, object detection scores. Some of these cues are computed at various relative locations of the provided human bounding box, yielding a total of 20 kernels. We evaluated the descriptors on the default and on the deformed part configurations. We optimize the weights for kernel combination using randomized grid search.

Experiments on the challenging PASCAL VOC 2012 dataset show that our method outperforms the state-of-the-art on the majority of action classes.

- [1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9):1627–1645, 2010.
- [2] L. Ladický, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *Proc. CVPR*, 2013.