Action Recognition From Weak Alignment of Body Parts

Minh Hoai[12]
http://www.robots.ox.ac.uk/~minhhoai/
L'ubor Ladický[3]
http://www.inf.ethz.ch/personal/ladickyl/
Andrew Zisserman[1]
http://www.robots.ox.ac.uk/~az/
[3]

[1] Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

[2] Stony Brook University
Stony Brook, NY, USA

[3] ETH Zürich
Zürich, Switzerland

Abstract

We propose a method for human action recognition from still images that uses the silhouette and the upper body as a proxy for the pose of the person, and also to guide alignment between instances for the purpose of computing registered feature descriptors. Our contributions include an efficient algorithm, formulated as an energy minimization, for using the silhouette to align body parts between imaged human samples. The descriptors computed over the aligned body parts are incorporated in a multiple kernel framework to learn a classifier for each action class. 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 Introduction

The objective of this paper is to recognize human actions in still images. This area has seen increasing interest over the last five years [3, 8, 12, 17, 18, 20, 22, 24, 26, 27]. Action recognition in still images is useful in its own right as a contribution to image understanding. It can also play a part in recognizing actions in videos [16]. A number of datasets and challenges have been released for this task, in particular the PASCAL VOC 'Action Classification Competition' [6] which includes such actions as reading and jumping.

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.

Previous authors have recognized the importance of human body parts and alignment. Human body parts can be inferred from human pose and the significance of human pose in indicating an action has been employed in [4, 24, 27]. However these works often suffer from

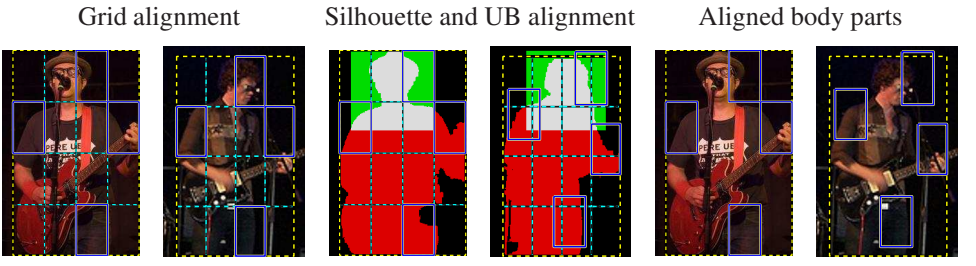


Figure 1: **Aligning body parts for action recognition**. (a) & (b): the alignment induced by a regular grid is not suited for registering body parts (c.f., solid blue boxes). The geometric constraints provided by the silhouettes and upper bodies ((c) & (d)) lead to a good alignment of parts. (e) & (f): 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).

incorrect pose estimation due to the challenges of severe occlusion and significant intra-class variability.

The importance of alignment has a long tradition in category recognition. For example, Berg *et al.* [1] proposed a method for recognizing objects using low distortion correspondences. Duchenne *et al.* [5] used a dense deformation field for aligning images. Ladicky *et al.* [14] proposed a locally affine deformation field for human detection. Bourdev *et al.* [2] used poselets for attribute classification. In the context of action recognition, Delaitre *et al.* [3] use the displacement of flexible parts in a Deformable Part Model (DPM) [7] to partially align training samples. However, as will be shown, the default variability of the DPM is insufficient for this task, while the silhouette and upper-body provide sufficient guidance for alignment, as illustrated in Fig. 1.

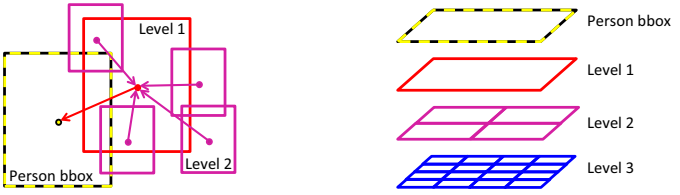
The rest of this paper describes our model for action recognition. Our classifier consists of several SVM kernels, whose combination is learnt for each action class. The kernels are constructed based on aligned images, where the alignment is guided by the silhouette and upper body (UB) detection. Throughout the paper, we illustrate our method using results and images from the PASCAL VOC dataset.

2 Weak alignment of body parts

This section describes a framework for alignment between human images using silhouettes and upper bodies. We also sketch, how the part alignment is used to obtain a feature vector. We propose to use silhouettes and the upper bodies instead of human poses, because pose estimation is less reliable and therefore ineffective (see Sec. 4.4). The algorithms to obtain silhouettes and detect upper bodies are described in Sec. 3.

2.1 Computing the alignment between two people

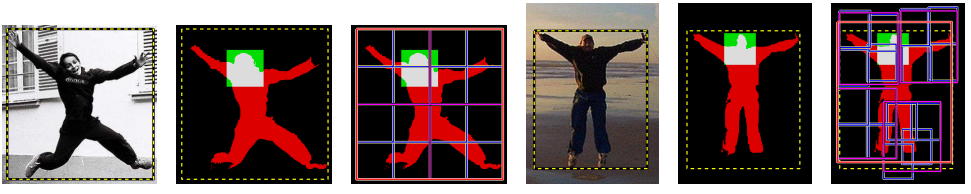
Given the bounding box of a human, we approximate the human body by a set of deformable rectangular parts, which is similar to a DPM [7]. The goal is to align these rectangular parts between two images, referred to as *reference* and *probe* images. We formulate the problem



(a) Spatial model for parts (b) Default part configuration

Figure 2: **Part model**. (a): parts are specified relative to its parent; the root part is anchored on the provided human bounding box. (b) default configuration of parts for the 3-level model.

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.



(a) ref. image (b) seg & UB (c) ref. parts (d) probe (e) seg & UB (f) aligned

Figure 3: **Aligning a probe image (d) to a reference image (a)**. First, we detect the upper body and perform human segmentation ((b), (e)). Second, we translate the parts of the probe image from the default locations to new locations (f) to be more similar to the parts in the reference image (c). This yields better alignment between two images, e.g., consider the 16 smallest parts (blue rectangles in the 4×4 grid), the bottom right in both images refers to the foot, the part at row 2 and column 2 is the underarm corner.

In more detail, 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 $\mathbf{p}_1^{ref}, \cdots \mathbf{p}_k^{ref}$ 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 $\mathbf{p}_1, \cdots, \mathbf{p}_k$ 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 (parts arranged in regular grids at multiple layers as in Fig. 2(b)). The

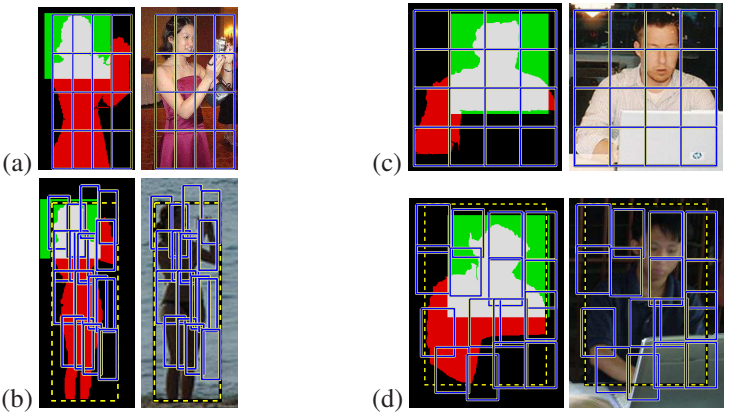


Figure 4: **Alignment examples**. (b) and (d) are aligned to (a) and (c), respectively. For visibility, only the level-3 parts are shown.

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).

The process of aligning one person w.r.t. another is illustrated in Fig. 3. Two other examples are given in Fig. 4.

Implementation details. The parts are grouped into layers, and the relative locations of parts are specified by a spatial model. Specifically, the i^{th} layer contains 4^{i-1} parts of the same size. Each part in the i^{th} layer is the parent of four parts in the $(i+1)^{th}$ layer, and the size of a parent part is twice the size of its children. The spatial model specifies the location of each part w.r.t. its parent with the exception of the root part being defined w.r.t. the bounding box of the human subject (Fig. 2(a)). Given the bounding box of a human subject, the default configuration is a set of layers of parts in regular grids, as illustrated in Fig. 2(b).

The inference for finding the configuration of parts that minimizes the above energy is efficient. First, this model only considers translational deformation. Second, computing the feature vectors for a particular part at all locations in an image can be done efficiently using an integral image. Third, the spatial model that relates the parts together is of a tree structure. This enables the use of dynamic programming and generalized distance transforms to efficiently search over all possible part configurations in an image, without restricting the possible locations for each part. This inference algorithm resembles the ones in [7, 25], but is simpler, because there is no mixture-of-components or mixture-of-parts. This inference algorithm scales linearly with the size of the probe image. For a 180×140 probe image, our implementation takes 360ms, running on a 2.3GHz Intel Core i7 machine.

2.2 Feature descriptor from multiple alignments

The previous sub-section describes how to align two people using their silhouettes and upper bodies. Because silhouettes and upper bodies are coarse approximation of poses, the above procedure works the best for pairs of similar images. Therefore, we propose to use the nearest neighbors to define the deformation space of an image and average the descriptors

computed at the deformed configurations.

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 spread of the deformation space depends on the similarity of its nearest neighbors. If all nearest neighbors are the same, a single deformation of the probe image attains better alignment. On the contrary, if every nearest neighbor is different in their own way, no single deformation suffices.

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 deformed configuration of parts. We pool the HOGs for each corresponding part by averaging. The process can be thought as alignment-informed jittering. Further details are explained in Sec. 4.2.

3 Silhouette and Upper Body Detection

3.1 Human silhouettes from segmentation

Human silhouettes are obtained using a foreground/background segmentation algorithm. This algorithm is based on a joint energy minimization framework [15] that consists of energy potentials from a pose model [25], a color model [21], and texture classifiers [13].

For a test image with a provided human bounding box, the silhouette extraction method proceeds in several stages. First, a large set of pose candidates is estimated using the mixture-of-parts model [25] with an additional hidden-part label [15]. These pose candidates are ranked by the pose detection score and the amount of overlapping between the pose's enclosing bounding box and the given human bounding box. For each of the top pose candidates, we apply the joints-to-segmentation mapping and train a foreground/background GMM color model [21] for the pixels which are mapped to. Using an energy minimization framework [15], we seek a foreground/background segmentation that yields the highest agreement with the candidate pose, the color probabilities, and the texture potentials. Examples of estimated poses and segmentations are given in Fig. 5.

In more detail, the pose model [25] is trained using 877 images from the VOC segmentation dataset containing 1532 person instances. In order to predict joint positions, all joint positions are manually annotated for a subset of these images. The joints-to-segmentation mapping is implemented as a classifier for pixels: the features consist of the distances of each training pixel to the ground truth locations of joints and limbs, and the classifier is learnt by boosting from the segmentations and joint positions [15]. The texture classifiers [13] are trained using VOC segmentation dataset which consists of 2913 segmented images. The texture classifiers use TextonBoost [23] and superpixels bag-of-words potentials [13].

Figure 5: **Pose and segmentation results**. Even though pose estimation is imperfect, it provides a good initialization for the foreground/background segmentation.

Due to the lack of ground truth annotation, a quantitative evaluation of the silhouette extraction algorithm on VOC action dataset is not possible. Qualitatively, it works better for images where a large part of the body is visible. For images where the person is in a sitting pose and severely occluded (e.g., reading, using computer), pose estimation may fail to provide a good pose candidate, and the segmentation result is consequently less reliable.

3.2 Upper-body detection

To localize the upper body, the Calvin upper-body detector[1] is used (this was the best upper-body detector, when this work was being developed; better detectors are now available [10, 11]). Unfortunately, the detection with the highest score does not always correspond to the actual upper body, perhaps due to the wide range of poses and scales present in the dataset. To reduce the number of false positives, we perform several post-processing steps. First, we remove detections that are either too big or too small. Second, we discard detections at improbable locations, such as detections at the bottom of the full body bounding box. For the remaining detections, if any, we dismiss all, but the one with highest detection score.

After post-processing, around 22% of the images have no upper body detection. Many of them correspond to a human subject at small scale with low contrast, and others correspond to images of close-up upper bodies. The former is more likely to happen for images of full body in running, walking, standing, or riding poses. For a full body in these poses, the height is often larger than the width, and a reasonable guess for the location of the upper body is the upper square of the full body bounding box. For close-ups of upper bodies, the middle square is often a reasonable guess for the location of the upper body. These observations lead to two heuristics for the location of the upper body in the case that the detector fails to return a detection. Fig. 6 illustrates these heuristics.

4 Experiments

This section describes experiments on the *Action dataset* from the PASCAL VOC2012 Challenge [6]. The performance measure is average precision, which is the standard measurement used by PASCAL VOC Challenge.

4.1 VOC2012 Action dataset

The Action dataset is the dataset for the "Action Recognition from Still Images" challenge. This dataset contains 11 action classes: jumping, phoning, reading, playing instrument, rid-

[1] http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/

Figure 6: **Heuristics for the upper body localization**, in case the detector fails. We use the top square for human subjects of which the height is larger than the width, and the middle square otherwise. In each image, the dash yellow rectangle is the provided bounding box of the human subject, and the solid cyan rectangle is the upper body obtained using these heuristics.

ding bike, ridding horse, running, taking photograph, using computer, walking, and others (images that do not belong to any of the first 10 classes).

The Action dataset consists of three disjoint subsets, voc-train, voc-val, and voc-test for training, validation, and testing respectively. The annotations of voc-test is not available to us, and the performance on this subset can only be obtained by submitting the results to the PASCAL VOC evaluation server. Because the evaluation server limits the number of submissions, it is impossible to use voc-test to report detailed analysis of our method and its components. To bypass this issue, we divide the validation data, voc-val, into two disjoints subsets, voc-val-1 and voc-val-2. The former consists of all images from VOC2011 Challenge, and the latter only contains images that are exclusive to VOC2012. voc-val-1 and voc-val-2 are used for tuning and testing respectively; the results reported in Sec. 4.3 are computed on voc-val-2. We only run our method on voc-test once, and the results returned by the PASCAL VOC evaluation server are used to compare with other competition entries. The number of human subjects (*not* images) in voc-train, voc-val-1, voc-val-2 and voc-test, are 3134, 1676, 1468, and 6283 respectively. Note the ROI (bounding box) is provided for each person.

4.2 Kernel SVM

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 optimize the weights for kernel combination using randomized grid search.

4.2.1 From feature descriptors to kernels

We compute a kernel for each type of feature descriptors. We investigated five different types of kernels: linear, Hellinger, intersection, RBF and RBF-$\mathcal{X}^2$. Our initial experiments showed that a RBF-$\mathcal{X}^2$ kernel outperformed other kernels for histogram-based features (SIFT and color), while a RBF kernel worked best for the remaining feature types. Consequently, we only report results for these kernels. The RBF-$\mathcal{X}^2$ and RBF kernels are defined as: $k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{\gamma}\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}\right)$ and $k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{\sigma}\sum_i (x_i-y_i)^2\right)$ respectively. In our experiments, γ and σ are set to be the average value of $\sum_i \frac{(x_i-y_i)^2}{x_i+y_i}$ and $\sum_i (x_i-y_i)^2$ respectively; the average is computed over all pairs of training examples $(\mathbf{x},\mathbf{y})$.

	jump	phone	instr	read	bike	horse	run	photo	comp	walk	mAP
DF	54.6	26.5	47.4	31.2	78.8	72.5	70.5	21.8	41.8	59.4	50.5
PA	**59.5**	**27.8**	**50.1**	31.4	**80.4**	**75.8**	70.9	21.5	**45.6**	58.9	**52.2**
Jitter1	52.6	25.3	50.9	28.4	78.7	72.1	69.0	18.6	39.6	58.2	49.3
Jitter2	53.1	22.1	48.8	30.9	75.1	64.0	63.1	18.1	40.3	51.6	46.7

Table 1: **Average precision for HOG-based kernels on the person bounding box**. DF: default kernel, PA: parts-aligned kernel. For each class, the larger value which is at least 1% better than the other is printed in bold. The last two rows are results of random jittering.

4.2.2 Parts-aligned kernels

We compare the performance of *default kernels* and *parts-aligned kernels*. A default kernel is the kernel computed for the feature descriptor extracted at the default configuration of parts. A parts-aligned kernel is a combination of the default kernel and the deformed kernel. The deformed kernel is the kernel constructed for the improved descriptor (computed from the average of the descriptors extracted at the deformed part configurations, as explained in Sec. 2.2), i.e. the improvement originates from a better alignment of body parts. This alignment process bridges the pose gap, but may also reduce the inter-class differences (e.g., between running and walking), making the classification problem harder. As such, the performance of the deformed kernel can be worse than the performance of the default kernel for classes that require discriminative poses. To achieve the best of both worlds, we combine the two kernels by averaging them, producing the so-called parts-aligned kernel.

4.3 Default kernels versus parts-aligned kernels

This section describes several experiments that compare the performance of parts-aligned kernels and default kernels (explained in Sec. 4.2.2), which will be referred as PA and DF, respectively. Here, we use a 3-level part model as illustrated in Fig. 2(b).

Tab. 1 displays the average precision for several different methods that extract HOG descriptors for the person bounding box (but at different locations of parts). As can be seen, PA significantly improves over DF for 6 out of 10 classes, and it performs competitively on the other classes. We also experiment with two methods that are based on random jittering. *Jitter1* is similar to PA, but with random deformations instead of nearest neighbor deformations. *Jitter2* is the method where each training example induces another training example using a random translation. Random jittering, however, shows no advantage over DF.

Tab. 2 reports the performance for several types of visual descriptors. In all cases, PA significantly outperforms DF. The advantageous gap is smaller for SIFT-based kernels. This is perhaps because SIFT and BoW representation are designed to tolerate misalignment errors.

4.4 Using noisy pose information

The previous subsection shows the empirical benefits of using silhouettes and upper bodies for human action recognition. An alternative approach is to use poses, but current outputs of pose estimation are noisy and therefore not directly useful for action recognition. For example, the default kernel with SIFT descriptors yield mAP of 49.8%. If we combine this kernel with noisy pose information, the mAP decreases to 47.9%. Thus, noisy pose information hurts the performance.

	color		silhouette		SIFT		SIFT+silhouette	
	DF	PA	DF	PA	DF	PA	DF	PA
mAP	27.3	**29.7**	25.9	**28.8**	49.8	**50.5**	46.8	**47.5**

Table 2: **Mean average precision (mAP) for several feature types**. *color*: based on the histogram of normalized-rg values. *silhouette*: based on the occupancy of silhouette pixels. *SIFT*: BoW representation. *SIFT+silhouette*: BoW representation, but discarding SIFT descriptors computed outside the silhouette.

	Stanford & MIT	Shenzhen Univ.	Hacettepe & Bilkent Univ.	Oquab et al. [19]	Ours
jumping	75.7	73.8	59.4	78.4	**79.6**
phoning	44.8	45.0	39.6	46.0	**49.5**
play'instru	66.6	62.8	56.5	**75.6**	67.5
reading	44.4	41.4	34.4	**45.3**	39.1
ridingbike	93.2	93.0	75.7	93.5	**94.3**
ridinghorse	94.2	93.4	80.2	95.0	**96.0**
running	87.6	87.8	74.3	86.5	**89.2**
takingphoto	38.4	35.0	27.6	**49.3**	44.5
usingcomp	**70.6**	64.7	55.2	66.7	69.0
walking	75.6	73.5	56.6	69.5	**75.9**
mean	69.1	67.0	56.0	70.2	**70.5**

Table 3: **Comparison with the state-of-the-art methods.**. The first three methods are entries from the VOC2012 Challenge. Oquab *et al.* [19] is a method that uses deep learning and exploits huge amount of ImageNet data. Best results are printed in bold. Our method is the new state-of-the-art for 6 out of 10 classes, and it performs best overall.

4.5 Comparison to the state-of-the-art

We compare the performance of our method with the results of PASCAL VOC2012 Challenge and the recent work of Oquab *et al.* [19]. For this experiment, we combine all kernels and train the classifiers using both training and validation data. We further double the amount of training data by left-right mirroring.

The results on the test set are obtained by submitting the output of our algorithm to the PASCAL evaluation server. This submission is done once; the results on the test set are shown in Tab. 3. Our method won the PASCAL VOC Challenge 2012. It achieved the best performance overall, and outperforms all other methods for 6 classes.

5 Discussion

We have proposed a novel framework for the alignment of deformable parts with an efficient inference algorithm. We have demonstrated the benefits of this approach for recognizing human actions in still images using the silhouette and the upper body. As the field of computer vision progresses, more and more body parts will be detected reliably. Those body parts can

be effortlessly integrated into our framework as they become available. Furthermore, we envision the benefits of incorporating object detectors in our framework, because the presence, location, and scale of interacting objects can reveal information about the pose. Like body part detectors, object detectors can be included in our framework with ease.

Acknowledgements

This work was performed while Minh Hoai and L'ubor Ladický were at the Visual Geometry Group, Department of Engineering Science, University of Oxford. This work was supported by EPSRC grant EP/I012001/1 and ERC grant VisRec no. 228180.

References

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[2] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the International Conference on Computer Vision*, 2011.

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference*, 2010.

[4] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Proceedings of the European Conference on Computer Vision*, 2012.

[5] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *Proceedings of the International Conference on Computer Vision*, 2011.

[6] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. www.pascal-network.org/challenges/VOC/voc2012/workshop/, 2012.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[8] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[9] M. Hoai. Regularized max pooling for image categorization. In *Proceedings of the British Machine Vision Conference*, 2014.

[10] M. Hoai and A. Zisserman. Discriminative sub-categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[11] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[12] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *Proceedings of the International Conference on Pattern Recognition*, 2008.

[13] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *Proceedings of the International Conference on Computer Vision*, 2009.

[14] L. Ladicky, P. H. Torr, and A. Zisserman. Latent SVMs for human detection with a locally affine deformation field. In *Proceedings of the British Machine Vision Conference*, 2012.

[15] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[16] I. Laptev and P. Perez. Retrieving actions in movies. In *Proceedings of the International Conference on Computer Vision*, 2007.

[17] L. J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proceedings of the International Conference on Computer Vision*, 2007.

[18] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[20] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, 2012.

[21] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[22] F. Sener, C. Bas, and N. Ikizler-Cinbis. On recognizing actions in still images via multiple features. In *Proceedings of the European Conference on Computer Vision*, 2012.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision*, 2006.

[24] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[25] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[26] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[27] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5D graph matching. In *Proceedings of the European Conference on Computer Vision*, 2012.