Example-Based Human Pose Recovery under Predicted Partial Occlusions

Ronald Poppe

Abstract. For human pose recovery, the presence of occlusions due to objects or other persons in the scene remains a difficult problem to cope with. However, recent advances in the area of human detection allow for simultaneous segmentation of humans and the prediction of occluded regions. In this chapter, we present an example-based pose recovery approach where this information is used. We effectively used the grid-based nature of histograms of oriented gradients descriptors to ignore part of the image observation space. This allowed us to recover poses directly, even in the presence of significant occlusions. We evaluated our approach on the HumanEva-I dataset, where we simulated different occlusion conditions. Without occlusion, we obtained relative 3D errors of approximately 69 mm. Our results showed approximately 10% increase in error when 20% of the observation is occluded. When 33% of the observation is occluded, the error is on average 15% higher compared to the observations without occlusions. These results showed that poses can be recovered from partially occluded observations, with a moderate increase in error. To the best of our knowledge, our approach is the first to investigate the effect of partial occlusions in a direct matching approach. Future work is aimed at combining our work with human detection.

1 Introduction

Human body pose recovery, or pose estimation, is the process of estimating the configuration of body parts from sensor input. When poses are estimated over time, the term human motion analysis is used. Traditionally, motion capture systems require that markers are attached to the body. These systems have some major drawbacks as they are obtrusive, expensive, and impractical in applications in which the observed humans are not necessarily cooperative. As such, many applications, especially in

Ronald Poppe
Human Media Interaction Group, University of Twente, Enschede, The Netherlands
e-mail: poppe@ewi.utwente.nl

R. Babuska & F.C.A. Groen (Eds.): Interactive Collaborative Information Systems, SCI 281, pp. 327–354.
springerlink.com

surveillance and human–computer interaction (HCI), would benefit from a solution that is markerless. Vision-based motion capture systems attempt to provide such a solution using cameras as sensors. In this chapter, we focus on applications that require real-time, robust pose estimates.

In general, there are two main approaches in vision-based human motion analysis: model-based (or generative) and model-free (or discriminative) [57]. The former approach uses a parameterized human body model which, for a given model instantiation, is matched to the visual observation. Guided by the projection error, the pose estimate is refined until a (local) optimum is found. This method is flexible as many parameters such as limb lengths, viewpoint, and appearance can be modeled. The drawbacks are the computational costly steps of model projection and projection-to-observation matching. As the pose space is generally large (20–60 dimensions), global optimization is not very practical. Therefore, local optimization is often used, which presents the risk of getting stuck at a local optimum. When multiple observations in time are available, this issue can be solved by taking a sampling-based approach (e.g., Condensation [30]), where many particles are used to estimate the cost surface. Due to the high dimensionality, a sufficiently large number of particles are needed to obtain a good estimate. Each particle brings an additional computational cost of model projection and projection-to-observation matching. In practice, model-based approach are therefore not suitable for real-time applications. Moreover, initialization remains a difficult problem with these methods.

Discriminative approaches do not use a human body model, but approximate the mapping from image to pose directly. They can be applied in real time and automatically provide initialization, but are less flexible in terms of encoding of parameters compared to generative approaches. Variations such as visual appearance and viewpoint should be encoded either explicitly in the parameter space or implicitly in the image representation. In this research, we take a discriminative approach where we explicitly encode viewpoint, as changes in viewpoint have a large impact on the image representation. We require our image representation to implicitly encode lighting variations and variations in body dimensions and clothing.

The focus of this chapter is on robust invariant image representations. We use a variant of histograms of oriented gradients (HOG, [11]), a holistic image representation. Discriminative pose recovery approaches are either example-based or regression-based. Regression-based approaches allow for faster evaluation but their precise implementation and parameter setting (number of experts, regression function, learning of the regression, and gating functions) influences the performance. This is undesirable, because it is less intuitive to attribute the performance to image representation or regression approach. Therefore, we use an example-based approach instead. The n visually most similar examples (nearest neighbors) are selected and the final pose is estimated to be the weighted interpolation of the n corresponding poses. We discuss regression-based and example-based approaches in Section 2.3.

In realistic scenarios, partial occlusion of the human figure in the image due to other persons or objects in the environment is common. We present an approach to handle these partial occlusions, if these can be predicted from a foreground

segmentation process. We assume that the location and scale of a human figure can be detected from an image or video. When occlusions occur, we also require that these areas are labeled. Due to the importance of this preliminary step, we discuss the process of human detection in slightly more detail in Section 2.1.

The remainder of this chapter is structured as follows. Section 2 discusses literature in the field of human motion analysis, with an emphasis on discriminative work. In Section 3, we introduce our adaptation of HOG and the occlusion-sensitive example-based pose recovery approach. Results on the publicly available HumanEva benchmark dataset are presented in Section 4, and we conclude in Section 5.

2 Related Work on Human Motion Analysis

Human motion analysis from images or video comprises many topics. Here, we focus on human pose recovery in Section 2.2. We discuss discriminative approaches in Section 2.3 as these can be applied in real time. Since our approach assumes that a human subject is segmented in an image, we briefly discuss the state-of-the-art in human detection first in Section 2.1.

2.1 Human Detection

Human detection and pose recovery can be seen as complementary tasks. In the detection task, the aim is to generalize over different poses, whereas in the recovery task, one wants to discriminate between them. We advocate a separation of these tasks. This eliminates the need to perform human detection and pose recovery simultaneously, as this would require large amounts of pose-annotated training pairs, which are costly to obtain.

The detection of human subjects from images is an important first step in the analysis of human pose or action. Recently, there has been an increased interest in this topic (see [33] for an overview). In general, human detection methods are either holistic or part-based. A *holistic* approach considers the human body as a whole [11, 21]. In many cases, a retinoscopic representation is used, where the human is assumed to be centered within a defined region of interest (ROI). Human detection is performed by sliding the window over the image and performing binary classification at each location. Dalal and Triggs [11] train a support vector machine on positive and negative examples, encoded as HOGs. All training examples are retained in [21], where Chamfer matching between a large set of pedestrian examples is performed hierarchically. Dong et al. [12] explicitly take into account inter-human occlusion. They extract foreground blobs of a single person, or a group of people. An example-based approach, assuming that for each blob the corresponding number of persons with their exact locations are annotated, is used to segment each person individually. Earlier work by Elgammal and Davis [14] used known color distributions of each person to segment persons under occlusion. The advantage of holistic approaches is that they generate relatively few false positives since much

information about the human body can be incorporated. The main drawback of a holistic approach is that occlusions cannot be dealt with without closer inspection of the scene.

In contrast, *part-based* approaches divide the human body into several parts, each of which can be modeled individually. Human detection is performed by looking at assemblies of these individual parts (e.g. [43, 44, 80]). Mohan et al. [44] find the head, legs, and the separate arms in a window by applying component-based classifiers. Then, an SVM over the individual part detectors is used to classify the entire window as human or non-human. The work of Mikolajczyk et al. [43] is similar in nature, but the focus is on the face and shoulders, which are encoded for frontal and side views separately. Felzenszwalb et al. [17] describe a person with a deformable part model, where each part is discriminatively learned from HOG descriptors. Niebles et al. [47] use a similar approach, but reduce the search space by applying a holistic detector first and relying on temporal continuity. Recent work by Wu and Nevatia [80] takes into account occlusions between persons in the scene. Such an approach is not only able to detect persons but can also determine which part of the observation is occluded. In recent work [81], they extend their approach to output pixel-level segmentations. Lin et al. [37] address the problem of finding suitable assemblies by introducing a tree of parts. Using re-evaluation, their method is also able to segment occluded persons from an image. They extend their approach in [36] to better handle variation in pose.

In general, part-based approaches generate many false positives for individual body segments. This can be explained since a body segment alone is often less discriminative compared to a full body. However, part-based approaches have a number of advantages over holistic methods. First, geometric constraints can be encoded efficiently. Second, by learning the detectors for parts individually, the combinatorial problem is effectively decomposed. Therefore, fewer training data of body-part templates is needed. Third, given a sensible assembly algorithm, humans can still be detected and segmented from the image even if parts are missing. This allows part-based methods to cope with partial occlusions from the environment or other persons.

Summarized, recent work has greatly advanced the quality of human detection. Especially the successful combination of detection and segmentation leaves us to believe that it is realistic to assume that a separation into foreground, background, and occluded area can be made. In the remainder of this chapter, we assume that such a segmentation is available.

2.2 Human Pose Recovery

The aim of human pose recovery is to find a numerical solution for a parameter estimation problem. The parameters that need to be recovered include not only those that describe the human body configuration but also viewpoint, body dimensions, body appearance, and clothing description can be part of the parameter space. Usually, we are only interested in the body configuration, and want to generalize over

body dimensions, appearance, and clothing and environmental factors such as illumination. Still, these have an effect on the observation. Dealing with those variations is one of the challenges of human pose recovery. Usually, an image representation is used that is partly invariant to illumination and clothing, for example, by considering silhouettes. Another challenge is to deal with the high dimensionality of the pose (or rather parameter) space, especially when real-time performance in needed. A global search for the best parameter combination is computationally infeasible.

As mentioned before, human pose recovery approaches are either generative (model-based) or discriminative (model-free). We argue that the high computational cost of the projection-to-observation matching of generative approaches makes them unsuitable for real-time applications. In this chapter, we therefore take a discriminative approach. In the next section, we describe literature that reports on discriminative work.

2.3 Discriminative Approaches to Pose Recovery

If no explicit human body model is available, a direct relation between image observation and pose must be established. In practice, this means that the image representation must generalize over variations in body dimensions, appearance, and clothing. In general, a far-off view is assumed, where perspective effects are negligible. When multiple cameras are employed, calibration is assumed. Discriminative algorithms automatically perform (re)initialization and can be used to initialize model-based approaches (e.g. [25, 65, 68]). Several works first use a discriminative approach to find certain key poses in time, and use a model-based algorithm to recover the poses in the intermediary frames (e.g. [18, 61]).

The training data must account for those parameters that we wish to recover, usually the pose representation and the viewpoint. Not all kinematically possible poses are also likely, and the training data implicitly forms a manifold in pose space. Due to the high nonlinearity of this manifold, the pose space should be covered densely to obtain faithful mappings. Dimensionality reduction can be used in pose or image space to facilitate the learning of the mapping as in [13, 15].

Two main classes of pose estimation approach can be identified: example-based (Section 2.3.1) and learning-based (Section 2.3.2). Example-based approaches retain all image–pose training examples. For a given input image, a similarity search is performed, and candidate poses are interpolated to obtain the pose estimate. Learning-based approaches avoid having to store a large amount of examples and approximate the mapping from image to pose space functionally by training on image–pose pairs.

2.3.1 Example-Based

Example-based approaches use a database of examples that describe poses in both image space and pose space. While no mapping from image to pose space has to be learned, the drawback of example-based approaches is the large amount of space

that is needed to store the database. Moreover, matching can be computationally costly, depending on the search scheme that is used.

In its simplest form, example-based approaches encode both the image part of the database and the observation into image representations and perform a linear search to obtain the closest matches, the nearest neighbors. The associated poses of these matches can be interpolated to allow for a more continuous range of pose estimates. Poppe [56] uses HOG representations, which encode edges while allowing for small variations in spatial arrangement. Results are presented from monocular and multi-view settings. In the multi-view case, the camera arrangement in training and test conditions is required to be the same. Silhouettes described using turning angle and Chamfer distance are considered by Howe [26]. In later work [27], optical flow information is used in addition. Fathi and Mori [16] only use motion information, which is invariant to illumination and texture.

When multiple synchronized cameras are available, a visual hull can be constructed. Van den Bergh et al. [3] approximate this hull using 3D haarlets, an extension to 3D of the haarlets proposed in [79]. They focus on pose recognition and learn a discriminative set of haarlets to maximize recognition performance.

Instead of using a direct distance measure, Sullivan and Carlsson [72] use deformation cost between examples and an input image. To improve the robustness of the point transferral, the spatial relationship of the body points and color information is exploited. Mori and Malik [45] employ shape contexts to encode edges. In an estimation step, the stored example are deformed to match the image observation. In this deformation, the location of the hand-labeled 2D locations of joints also changes. The most likely 2D joint estimate is found by enforcing 2D image distance consistency between body parts.

Temporal information can be used to overcome ambiguities from the image to some extent. Toyama and Blake [76] incorporate examples in a probabilistic temporal framework. By employing an HMM, they approximate a low-dimensional manifold by linear segments. Similar in concept is the work by Ong et al. [52], who cluster the examples and determine flow vectors for each cluster. A particle filter framework is used where the particles are guided by the flow vectors. Particle likelihoods are based on the matching distance to the closest example in the cluster.

The computational complexity of a naive nearest neighbor search is linear in the number of examples. For recovering more unconstrained movements or high number of DOF, the number of required examples grows substantially. Therefore, Shakhnarovich et al. [67] introduce parameter sensitive hashing (PSH) to rapidly estimate the pose given a new image. Because of the ambiguity in the use of silhouettes alone, they use edge direction histograms within a contour. An alternative approach to reduce the computational complexity of the matching is by storing the examples in a tree (e.g. [21]). Given an input image, a top-down matching procedure is used. Starting from the highest level node, a matching is performed for each of the child nodes. Only those subtrees that satisfy a certain criterium (e.g., threshold or best match) are further evaluated. This significantly reduces the computation time needed to select similar examples. This approach has also been taken by Yang and Lee [83], who construct the pose estimate as a linear combination of the selected

examples from the bottom level of the tree. Rogez et al. [63] use a collection of trees. The nodes in each tree are trained to be discriminative and take into account a single dimension from a HOG representation. By using a collection of trees, many features can be used, and the resulting algorithm is more robust to noise.

2.3.2 Learning-Based

Learning-based approaches approximate the mapping from image to pose space functionally. The advantage of these regression methods is that inference can be performed efficiently and training data can be discarded after training. The drawback is learning the mapping. Especially for large amounts of training examples, computation requirements might be prohibitively large.

Xu and Hogg [82] present one of the earliest uses of regression in human pose recovery. A neural network is employed to map silhouette representations to pose representations. Agarwal and Triggs [2] use nonlinear relevance vector machine (RVM) regression over both linear and kernel bases to model the relation between histograms of shape contexts and 3D poses. Ambiguities are resolved using dynamics. Agarwal and Triggs [1] use direct regression to recover upper-body poses. Non-negative matrix factorization (NMF) on grid-based edge histograms is used to obtain a set of basis vectors that correspond to local features on the human body and ignore the presence of clutter. This enables them to recover poses without relying on background segmentation. The work by Onishi et al. [53] is similar in spirit, and extends the work of Poppe [56] with a noise-reduction step. Instead of applying NMF, they perform PCA in each block of cells in the grid to reduce the influence of backgrounds.

Instead of using a single view, information from multiple synchronized views can be combined into a voxel model. Ambiguities caused by using a single view are thus avoided. Also, a 3D voxel representation is independent of the camera setup. The approach is most suitable for controlled settings, where clean silhouettes can be obtained. Sun et al. [73] use an adaptations of the RVM to recover the pose from 3D shape context descriptors. When rotation normalization can be performed, such an approach can be used to learn view-independent regressors. Gond et al. [22] fit the voxel model in a 3D circular grid. This descriptor is rotation-normalized after recovering the orientation of the torso. The normalized feature representation is then used as an input for a sparse regressor. The approach has the advantage that significantly less training data is required, at the cost of an additional normalization step.

The space of common human poses is much smaller than the space of kinematically possible poses and these poses usually occupy a well-defined area in this high-dimensional space. This has led to the introduction of dimensionality reduction techniques. These techniques are also well suited for learning-based approaches as they can simplify the regression functions. For example, Grauman et al. [23] describe a distribution over both multi-view silhouettes and 3D joint locations with a mixture of probabilistic PCA. A pose estimate is obtained from the Bayesian reconstruction given the image representation. Similar in concept is the work of Bowden et al. [8], who fit a nonlinear point distribution model (PDM) to 2D position of head

and hands, the 2D body contour, and the 3D pose representation. The feature space is projected on a lower dimensional space and allows for reconstruction of the pose given an input image. Ong and Gong [51] include views from multiple cameras in the PDM and recover a pose from multi-view images. Rogez et al. [62] use single view and learn separate models. Temporal and spatial constraints are further used to solve pose ambiguities. This concept is similar to Brand's [9], who models a manifold of pose and velocity configurations with an HMM. Temporal ambiguities are resolved by recovering poses over an entire sequence by applying the Viterbi algorithm.

Due to depth ambiguities in image space, the mapping from image to pose space is multi-valued and cannot be determined with a single regressor. Therefore, mixtures of regressors have been introduced. These divide the image space into clusters, where a regressor is learned for each cluster. Rosales and Sclaroff [64] cluster the 2D pose space and learn specialized functions for each cluster from image descriptors to pose space. A neural network is used as mapping function. In [66], the work is extended to allow input from multiple cameras. The pose is estimated for each camera individually and in a subsequent step, the hypotheses are combined into a set of self-consistent 3D pose hypotheses. Thayananthan et al. [74] use a mixture of regressors but validate the pose estimate for each by matching it against the input image to select the most likely pose. A similar approach is used by Sminchisescu et al. [70], who jointly learn mappings between image and pose space. The processes are guaranteed to converge to equilibrium. During inference, the results of the mapping from image to pose is validated using the mapping back.

Sminchisescu et al. [71] take a probabilistic approach and model the multi-valued nature of the mapping with Bayesian mixture of experts (BME). Each expert has an associated gating function, which gives the conditional probability that the regressor should be used given an input image. Guo and Qian [24] adapt the initialization using k-means and use stereo observations to reduce the multi-modality of the mapping. Ning et al. [48] initialize the experts on a partitioned subset of the image space. In the BME framework, experts and gating functions are learned simultaneously. This requires a double-loop optimization approach, which is computationally costly. Therefore, Bo et al. [6] train both models sequentially, which results in a decrease of both memory and computation requirements. Their algorithm thus can handle much larger numbers of examples.

Usually, not the whole image representation is useful for learning the regression. Redundancy and noise in the training data can thus affect the learning and performance of the regressors. This can be avoided by selecting only the relevant features. Additionally, this lowers the dimensionality of the image space and thus the complexity of the regressor. Ning et al. [49] jointly learn the BME regressors and the selection of visual words in a supervised manner. A similar approach is taken by Kanaujia et al. [31], who focus on hierarchical image representations and semi-supervised learning. Okada and Soatto [50] discriminatively select those orientations within HOG cells that are meaningful for predefined class of poses. An input image is first classified to a pose class, before recovering the pose. Bissacco

et al. [4] use boosting to select a limited set of discriminative binary edge features and to learn the mapping directly.

Instead of learning the regression function offline, Urtasun and Darrell [78] learn it online, given an input image. With an example-based approach, the closest examples are selected. A local regression is then learned from these matches. Their approach can handle large numbers of examples, but is computationally more costly due to the selection of the nearest neighbors.

Learning these mappings depends largely on the availability of pose–observation pairs, which are difficult to obtain, especially in more unconstrained scenarios. Several authors have used synthetic observations generated by character modeling software (e.g. [2, 70]). Instead, Navaratnam et al. [46] use unlabeled examples to improve the regression functions. These examples can be easily obtained by using images of humans and by considering motion capture data.

3 Pose Recovery Using HOG

To describe an image, we can either use a holistic descriptor or a local (or patch-based) descriptor. The former encodes the image observation as a whole. Local deformations in the image will affect the entire descriptor. In contrast, local descriptors describe the image observation as a collection of local regions. Usually, these regions are extracted at interest points (local features), which are expected to be invariant to changes in viewpoint and illumination [41, 77]. Currently, a popular local descriptor is the scale invariant feature transform (SIFT [39]) and extensions (SIFT-PCA [32] and GLOH [42]). Local descriptors have the advantage that they can cope with variations in illumination, pose, and viewpoint to some extent. However, they strongly rely on robust extraction of interest points, which might be difficult due to differences in subject and background appearance. Moreover, extraction of local descriptors is more time-consuming due to the localization of interest points and the calculation of the local descriptor. Also, matching of bags of local descriptors is less straightforward. Therefore, we use a holistic descriptor in our work.

In this section, we present an example-based approach to human pose recovery. We use HOG as image representation. This holistic representation has been introduced by Dalal and Triggs [11]. Their HOG descriptor is inspired by work on orientation histograms [19], but uses dense sampling instead. The key idea is to calculate HOG (edges) within each cell of a regular spatial grid. This grid has a fixed number of cells which cover an area that is determined by a rectangular ROI. The HOG descriptor is a concatenation of all cell histograms. Several alternatives to HOGs have been proposed in literature. Levi and Weiss [35] use edge orientation histograms that contain ratios between orientation responses, dominant orientation, and symmetry features, calculated exhaustively over all rectangular subwindows of an image. Adaboost is used to select the relevant features. In contrast, HOGs use a fixed spatial structure, which allows for direct matching. The pyramid of histograms of oriented gradients (PHOG) proposed by Bosch et al. [7] is a generalization of the HOG where the notion of a block of cells is extended to multiple levels. At the

lowest level, the ROI is described as a single edge orientation histogram. For each higher pyramid level, a division into 2×2 cells is made. The PHOG approach is suitable when there is variation in the localization of the ROI but restricts the number of rows and columns in the grid to be equal and to be powers of 2. As the height of a human figure in the image is larger than its width, we use the original HOG concept.

HOGs have been used for several human motion analysis tasks. Dalal and Triggs initially used HOGs for pedestrian detection, a binary classification task. Variations in clothing, lighting, and body dimensions, but also viewpoint and pose, were implicitly encoded. Such an approach is reasonable since there are clear cues such as head and shoulder lines, which remain present also when seen from different viewpoints. Gandhi and Trivedi [20] use HOG descriptors to classify the orientation of pedestrians, thus explicitly encoding the (relative) viewpoint. Thurau [75] used HOGs to model human shape for human action recognition. Both Liu et al. [38] and Chakraborty et al. [10] use body part classifiers based on HOG descriptors. Such an approach is suitable for 2D location of limbs. However, without strong pose priors, lifting these to 3D will lead to ambiguities as there is no verification step where the observation is used in a holistic manner.

While HOGs have been shown to be robust descriptors for the aforementioned tasks, we believe that HOG descriptors are even sufficiently rich for recovery of human poses, including the viewpoint. This task is, however, more demanding as we do not have to distinguish between a small number of classes, but instead aim at *regression* of 60-dimensional poses. Moreover, the HOG descriptors still have to be invariant to lighting, clothing, and body dimensions.

Since we need to recover more information from the HOG descriptors, we also require more precise HOG extraction. The above-mentioned works extract the HOG descriptors directly from the image, which has two drawbacks. First, the ROI needs to be determined, which is computationally expensive. The ROI can vary in position and scale (we do not regard rotation, upright recordings are assumed), and many possible ROI candidates within an image have to be validated. Zhu et al. [84] introduce an efficient approach based on the integral image [60], but real-time performance still cannot achieved. Moreover, there will be false positives in the neighborhood of the actual ROI, which makes determination of the actual location and scale difficult.

Second, there is the problem of background clutter. Edges within the ROI that do not belong to the person, but to the background, will affect the HOG descriptor. Therefore, a number of works have explored ways to focus on edges that belong to the foreground. Poppe [56] uses background subtraction and only uses those edge responses that fall within the foreground region. Agarwal and Triggs [1] use NMF to suppress background edges. They demonstrate their work on recovery of frontal poses. Both Sminchisescu et al. [70] and Okada and Soatto [50] implicitly determine a set of discriminative features by learning regression functions from the HOG space to the pose space. Due to the use of multiple regressors, this selection is pose-dependent. Rogez et al. [63] use randomized trees, each of which is trained

discriminatively. However, none of these works has explicitly addressed partial occlusions.

Given the common presence of partial occlusions due to other persons or the environment, there has been surprisingly little work that explicitly addresses this issue. Poppe and Poel [59] detect humans and recover their poses in single images. They use body-part templates and, for a match, vote over all joints in the human body. Such an approach can deal with severe occlusions, but is restricted to a limited class of motions (e.g., walking). Peursum et al. [55] use factored-state hierarchical HMM to model the motion of one given action. Occlusion of the feet can be detected using [54], and the likelihood function is adapted by ignoring the occluded area. The learned dynamical model will ensure stable tracking but also poses restrictions on the movement that can be recovered.

For example-based approaches, occlusions are variations that are not explicitly modeled in the example set. To be able to handle these variations, there must be a way to take into account the missing (or ambiguous) information in the matching. To the best of our knowledge, only one paper takes into account occlusion in an example-based approach. Howe [28] uses boundary fragment matching to match partial shapes. Boundary fragments are small parameterized outlines of an extracted silhouette. Background and occlusion areas need to be labeled, so the matching algorithm knows which boundary fragments belong to the actual shape.

In many cases, silhouettes can be obtained relatively reliably using background substraction. Similar to [56], we assume that such a segmentation into foreground and background can be made. Employing this segmentation has two main advantages. First, determination of the ROI is straightforward. This relieves the burden on the detection task, as only a single detection window has to be processed. This will significantly aid in achieving real-time performance. Second, by considering only foreground edges, we effectively ignore background clutter. The resulting HOG descriptor is therefore not dependent on the background, which increases generalization. In addition, we assume that it is known which parts of the human subject in the image are occluded, similar to [28]. We effectively use the grid-based nature of the HOG descriptor to use only those dimensions in the matching procedure that correspond to non-occluded cells. Our work is an adaptation of [56] where descriptor normalization and matching procedure are adapted. In this section, we explain the steps of our approach and show that poses can be recovered accurately, even when foreground segmentation is noisy. We further demonstrate that our approach can recover poses under partial occlusion. To the best of our knowledge, our approach is the first to address partial occlusions in a direct matching approach. We believe that this is a key characteristic of any human pose recovery approach that is to perform in real time in more realistic environments.

In our contribution, we neither regard the temporal aspect nor do we apply any measures to reduce the computational complexity. This allows us to focus on the performance of the HOG descriptors. In Section 3.1, we discuss our HOG variant, and how we obtain the descriptor from an image. The nearest neighbor pose recovery approach is explained in Section 3.2. Our experiments on the HumanEva datasets are presented in Section 4.

3.1 Histogram of Oriented Gradients

Our descriptor differs from the HOG descriptor as described by Dalal and Triggs [11]. First, we only take into account the edges within a foreground mask. This requires background segmentation, but allows us to focus only on those edges that are meaningful. Second, we do not use the notion of (overlapping) blocks, which results in a significantly reduced descriptor size. Third, we do not apply color normalization, which further reduces the computational costs of calculating the descriptor. Fourth, we use a different grid size. In our experiments (Section 4), we divide the ROI into a grid with 6 rows and 5 columns. This is an arbitrary choice, but the height of each cell roughly corresponds with the height of the head in a standing position. Similarly, in a relaxed standing pose, the body covers approximately 3 columns horizontally.

Within each cell in the grid, we calculate the orientation and magnitude of each pixel that appears in the foreground mask. We apply a $[-1\ 0\ 1]$ gradient filter to each pixel in horizontal and vertical direction independently. The orientation of the edge is given by the angle between these two derivatives. The magnitude is given by the square root of the sum of the two squared derivatives. We divide the absolute orientations over 9 equally sized bins in the 0°–180° range. Each pixel contributes the magnitude of its orientation to the according histogram bin, which results in a 9-bin histogram per cell. Note that this binning is slightly different than proposed in [11], where votes are interpolated bi-linearly between the neighboring cells and orientation bins. The total length of the descriptor is 270. Poppe [56] normalized the entire descriptor to unit length to overcome differences in scale. To further reduce the size of the HOG descriptor and suppress background noise, PCA is applied by Lu and Little [40] and Onishi et al. [53] in the context of human motion analysis. However, both a global normalization and the use of PCA make the descriptor holistic, as local variations affect the entire descriptor. However, partial occlusions cause some of the observation to be uninformative, or even misleading. Normalization of the entire descriptor depends on all individual cells. In case of occlusion, some of the cells will have different edge responses. By normalizing descriptor d to unit length, these cells will affect all others. Therefore, we normalize each cell $h_{i,j}$ (i and j are a row and column index, respectively) individually to be of unit length. This has the advantage that we can still deal with variations in scale, as each cell is approximately equal in size. On the other hand, we discard the global character, and each cell contributes equally to the final descriptor. Specifically, individual cells that have a relatively low edge response have a similar summed weight as cells that have a high response. Alternatively, we could have normalized each cell by its area size. This would make the descriptor invariant to scale, but would not take into account variations due to different lighting settings and clothing.

3.1.1 Determination of ROI and Foreground Mask

We calculate HOGs within an image's ROI, in our case, the bounding box around the subject. While HOGs can be used to determine this region, as in [11, 75, 84], we

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

Fig. 1 Different steps in the calculation of a HOG descriptor. (a) the input image, with simulated occlusion, (b) the foreground mask. For clarity, the occlusion is not shown. (c) the edge magnitudes and (d) the HOG descriptor. The cells that are (partly) occluded are shown black. (e) Locations of the 20 joints.

rely on background subtraction. As discussed previously, this significantly speeds up the process, and we suppress background edges at the same time. We describe the process here in detail to allow for replication. First, we apply the background subtraction with the suggested risk values, as included in the HumanEva source code [69]. The background is modeled as a mixture model with 3 Gaussians per color channel. The minimum enclosing box of all foreground areas larger than 600 pixels is obtained. After conversion to HSV color space, we apply shadow removal in the lower 20% of the ROI. Pixels that have a saturation that is between 0 and 25 higher than the saturation of any of the means in the background mixture model are removed from the foreground mask. We again obtain the minimum enclosing box, which is our final ROI. Figure 1 shows an example of background subtraction and displays the HOG descriptor.

It may seem that our approach is highly sensitive to good background subtraction, but the shadow removal is only needed to ensure that the ROI fits the subject reasonably. For certain cases, we slightly adjusted the parameters. For camera 1 in HumanEva-I, only for subject 2, we multiplied the risk with factor 10^{12} to remove artifacts from the foreground. For cameras 2 and 3 in HumanEva-I, we lowered the shadow threshold to 10. We did not use the additional four grayscale cameras in HumanEva-I. For HumanEva-II, we reduced the background risk with factor 10^{50}, only for camera 3. Still, errors in the background segmentation frequently result in incorrect determination of the ROI and inaccurate foreground masks (see also Figure 2).

(a) (b) (c) (d)

Fig. 2 Background subtraction errors that result in inaccurate foreground masks and incorrect placement of the ROI

3.2 Pose Recovery Using Nearest Neighbor Interpolation

In an example-based approach, each image observation is encoded and matched against an example set of encoded observations. We use the previously described HOGs as encodings. To match a HOG with those in the example set, we need to define a distance measure between the two descriptors. To deal with partial occlusions, we match descriptors at the cell level. We introduce weights $\phi_{i,j}$ for each cell. These weights scale the feature space and therefore affect the distance functions that we define. While these weights can take any value (e.g., in the $[0,1]$ range), we use here binary values. That is, $\phi_{i,j} = 0$ in the case of occlusion within the cell, and $\phi_{i,j} = 1$, otherwise. This weighting effectively determines a lower-dimensional subspace, in which we can perform matching. The distance between descriptor d with cell-normalized histograms $\hat{h}_{i,j}$ and descriptor d' with cell-normalized histograms $\hat{h'}_{i,j}$ is calculated as

$$D(d,d') = \frac{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c}\phi_{i,j}\delta(\hat{h}_{i,j},\hat{h}'_{i,j})}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c}\phi_{i,j}} \tag{1}$$

Here, n_r and n_c are the number of rows and columns, respectively. In our experiments, $n_r = 5$ and $n_c = 6$. δ is the distance measure between two histograms, for which we use Manhattan distance. Note that, since we predict which cells are partly occluded, we could ignore these cells and normalize the descriptor to unit length. This approach would maintain the advantages of a global normalization but has the drawback that normalization of all n examples in the database needs to be performed at run time. This will severely affect the computational performance.

Matching a HOG with the entire example set results in a distance value for each of the m examples. We could choose the example with the lowest distance, as this

is the example that best matches the image observation. However, in practice, taking the n best matches (nearest neighbors) results in more accurate pose recovery. It should be noted that n is the only parameter in our approach. Of course, n will depend on the number of examples in the example set that are close to the presented frame. Here, we use $n = 25$, in accordance with [58]. To determine the final pose estimate, we use the poses that correspond to the n best examples. We determine the final pose estimate $\boldsymbol{p}$, the normalized weighted interpolation of these poses, as $\boldsymbol{p} = \sum_{i=1..n} w^i \boldsymbol{p^i} / (\sum_{j=1..n} w^j + \delta)$ and $w^i = \frac{1}{d^i + \delta}$ $(1 \leq i \leq n)$. Here, δ is a small number to avoid division by zero in the rare case which, retrieved examples, exactly that match. $\boldsymbol{p^i}$, d^i, and w^i correspond to the pose vector, HOG distance value, and weight of the i^{th} best matching example, respectively. This implies that close HOG matches contribute more to the final estimate than HOG matches at a larger distance. One word of caution is in its place here. Since we interpolate poses, the final joint estimates are likely to lie closer to the mean distance for this joint, so closer to the body. This effect is especially visible for examples that have similar image observations but are distant in pose space.

Since we do not determine the correspondences between our localized subject in the image and the estimated pose, we are not able to estimate the global position of each joint. Instead, we report the distances of each joint relative to the pelvis (*torsoDistal*) joint, as suggested in [69] (see Figure 1(e)).

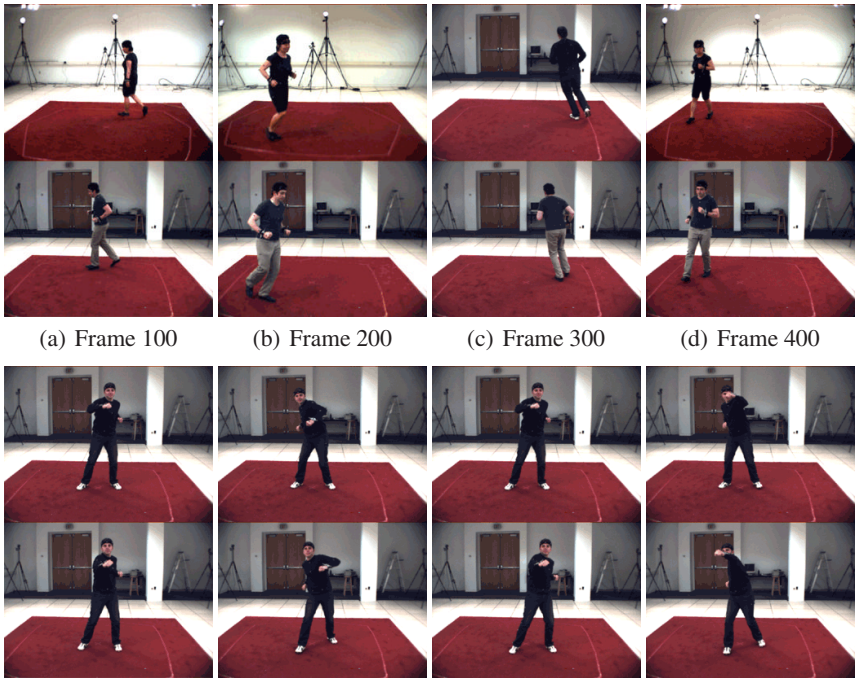
4 Experiment Results

In our experiments, we use the HumanEva-I dataset [69], which we describe in Section 4.1. We introduce the example set in Section 4.2. To the best of our knowledge, there is no dataset that contains both ground truth motion capture data and video data with occlusions. Therefore, we use the HumanEva dataset and simulate different types of occlusion. The test sets are presented in Section 4.3. Results on both non-occluded and occluded observations are presented in Section 4.4 and discussed in Section 4.5.

4.1 HumanEva Dataset

The HumanEva-I dataset [69] contains several sequences, divided into training, validation, and test sets.

The training and validation sequences in HumanEva-I contain synchronized video and motion capture (mocap) data. There are 4 subjects that perform 5 actions (Walking, Jog, Box, Gesture, and Throw/Catch). In addition, there is one sequence for each subject–action pair that contains only mocap data. In the test set of HumanEva-I, all these subject–action pairs also appear. Also, for each subject, there is an additional Combo sequence that contains walking, jogging movements, and some additional balancing movements that do not appear in any training or validation trial. Example frames are shown in Figure 3.



(a) Frame 100 (b) Frame 200 (c) Frame 300 (d) Frame 400

(e) Frame 100 (f) Frame 200 (g) Frame 300 (h) Frame 400

Fig. 3 Single best estimate (top row) and original frame (bottom row) for the HumanEva-I Jog sequence performed by subject 2 (a–d) and HumanEva-I Throw/Catch sequence performed by subject 3 (e–h), both evaluated using a single camera (C1)

The walking and jogging actions are performed by moving in a circle, counter-clockwise. Boxing, gesturing, and throwing and catching a ball are performed while facing camera C1. There is some variation in these sequences, though. Especially, the body orientation while catching the ball is heavily dependent on were the ball appears.

In HumanEva-I, each sequence has been recorded with 3 color cameras and 4 grayscale cameras. These have all been synchronized and calibrated. The frame rate of the video sequences is 60 frames per second. The dataset comes with source code to temporally align different video streams with the motion capture data. Also, background subtraction is included, which describes the background with a Gaussian mixture model. We have used these provided algorithms where possible. In our experiments, we use only the color cameras. For the monocular case, we only regard camera 1. Sequences of subject 4 were not evaluated due to difficulties with the background subtraction.

For the test sets, ground truth pose information is held back. An online validation system, as described in [69], is used to validate the pose recovery results. This system ensures that results of different parties can be compared, and frustrates

Table 1 Number of valid examples per action and subject in HumanEva-I training and validation sequences

Action	Subject 1	Subject 2	Subject 3	Total
Walking	1176	876	895	2947
Jog	439	795	831	2065
Throw/Catch	217	806	0	1023
Gestures	801	681	214	1696
Box	502	464	933	1899
Combo	0	0	0	0
Total	3135	3622	2873	9630

parameter tuning. Specifically, ground truth information in our case are the 3D positions of 20 key joints, relative to the pelvis (*torsoDistal*) joint (see Figure 1(e)). This is the full set of joints, and we report the root mean squared error (RMSE) in mm, averaged over all joints, as described in [69].

4.2 Example Sets

We describe our example set for monocular pose recovery. We associate the HOGs for an individual view with their corresponding poses. Only the examples that contain valid mocap data are included in the example set.

When given a new image observation, together with the knowledge from which camera the observation is obtained, we can estimate the relative pose. We observe that the elevation (rotation in vertical direction) and roll (rotation around line of sight) of all cameras are approximately the same. In other words, the orientation of all cameras is almost equal except for the orientation around a vertical axis. If we would rotate the subject in the scene around a vertical axis, we would theoretically be able to generate very similar observations for all cameras. In practice, view-specific parameters such as backgrounds and lighting conditions are likely to result in observations that are somewhat different. However, we want our approach to be robust against these image deformations and therefore we perform this rotation virtually. This has the additional advantage that the number of examples is effectively tripled, resulting in a total of 28,890. Table 1 summarizes the origins of the examples.

We transform the mocap data in such a way that we obtain the joint positions as if we were looking through another camera. With an observation from camera i, and the projection onto camera j, our pose vector $\boldsymbol{p_i} = (x_i, y_i, z_i, 1)^T$ is transformed into $\boldsymbol{p_j}$ as follows: $\mathbf{p_j} = M_j M_i^{-1} \boldsymbol{p_i}$, where M_i and M_j are the rotation matrices of cameras i and j, respectively.

4.3 Test Sets

We use the HumanEva-I test sequences of subjects 1–3 for testing. Evaluation of the Combo and Throw/Catch test sequences of subject 1 failed repeatedly.

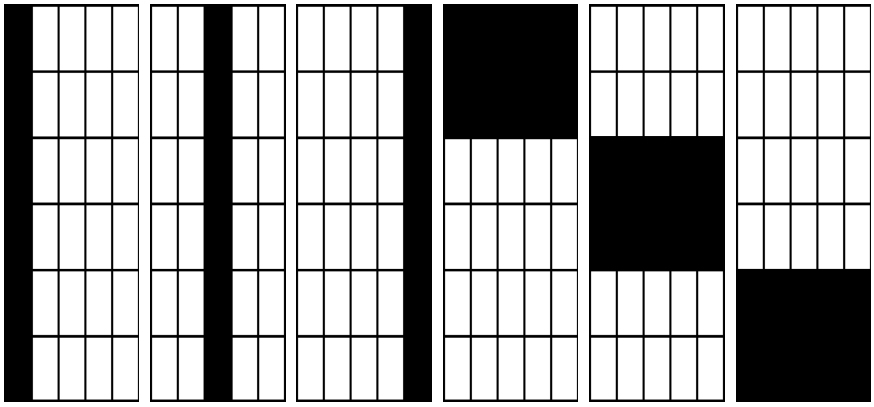


Fig. 4 Occlusion settings (left to right) *v_left*, *v_center*, *v_right*, *h_top*, *h_center*, and *h_bottom*

Consequently, we cannot report our results on these trials. Note that part of the Combo sequences contains movements that are not in the example set, which our algorithm cannot handle.

The original frames are used to evaluate the pose recovery accuracy without occlusion. In addition, we simulate different occlusion settings. We define six different occlusion settings. Each of these settings is a fixed combination of weights $\phi_{i,j}$. Effectively, we simulate occlusion on a fixed set of cells, not regarding the location of the subject in the image. As such, we can test the influence of occlusion on a large number of poses. The six settings that we define are *v_left*, *v_center*, *v_right*, *h_top*, *h_center*, and *h_bottom*, see also Figure 4. In the vertical and horizontal conditions, 20% and 33% of the image observation is occluded, respectively.

In addition to the fixed occlusion settings, we also created a *pole* sequence. This is the walking sequence of subject 1, but with a fixed occluded area in the image, not relative to the ROI. The area is a vertical pole with a width of 40 pixels. For comparison, the subject in the image is on average approximately 115 pixels in width. Due to the scale and pose, this can vary between 70 and 170 pixels. The *pole* sequence is a more realistic scenario and can show how the estimation error changes as the subject moves through an occlusion.

4.4 Results

Table 2 summarizes the average 3D errors over all joints, relative to the pelvis. The last column shows the results without occlusion. The average 3D relative error per joint is approximately 69 mm over all actions. For walking and jogging, the error is significantly lower at 45.05 mm and 45.98 mm, respectively. The larger number of available examples and lower variation in movement are the most important causes for the lower errors.

Table 2 Mean relative 3D error in *mm* per joint for HumanEva-I test sequences, evaluated with camera C1. For each fixed occlusion setting, the increase over the non-occluded observations and the amount of occlusion are given.

Subject	Action	Vertical			Horizontal			None
		Left	Center	Right	Top	Center	Bottom	
S1	Walking	39.80	43.28	42.10	47.52	43.83	45.83	39.03
S1	Jog	58.24	57.16	54.09	59.34	52.27	63.42	48.75
S1	Throw/Catch	N/A	N/A	N/A	N/A	N/A	N/A	N/A
S1	Gestures	30.72	28.93	30.75	30.58	30.44	31.79	30.11
S1	Box	84.40	93.13	84.80	97.81	90.30	79.41	81.38
S1	Combo	N/A	N/A	N/A	N/A	N/A	N/A	N/A
S2	Walking	37.60	40.18	39.22	42.23	39.53	41.24	37.27
S2	Jog	43.11	41.66	43.34	49.00	45.16	50.79	41.37
S2	Throw/Catch	73.53	72.13	71.34	76.81	74.06	76.33	72.18
S2	Gestures	95.30	95.65	86.14	84.90	85.25	93.16	82.81
S2	Box	109.17	120.03	107.43	107.14	112.11	115.45	105.08
S2	Combo	71.20	72.72	72.67	78.30	74.14	76.05	68.73
S3	Walking	58.38	64.73	64.23	61.31	61.22	65.79	58.86
S3	Jog	52.57	54.55	50.77	53.26	52.55	53.63	47.83
S3	Throw/Catch	120.06	112.22	94.38	110.81	103.15	122.22	101.40
S3	Gestures	80.09	76.85	97.63	72.49	91.20	114.00	82.63
S3	Box	105.64	110.48	98.69	110.15	105.65	106.99	98.44
S3	Combo	117.75	116.97	111.11	119.18	112.84	116.02	106.85
Average		73.60	75.04	71.79	74.43	73.36	78.26	68.92
Increase (%)		6.79	8.88	4.16	7.99	6.44	13.56	0.00
Occluded (%)		20.00	20.00	20.00	33.33	33.33	33.33	0.00

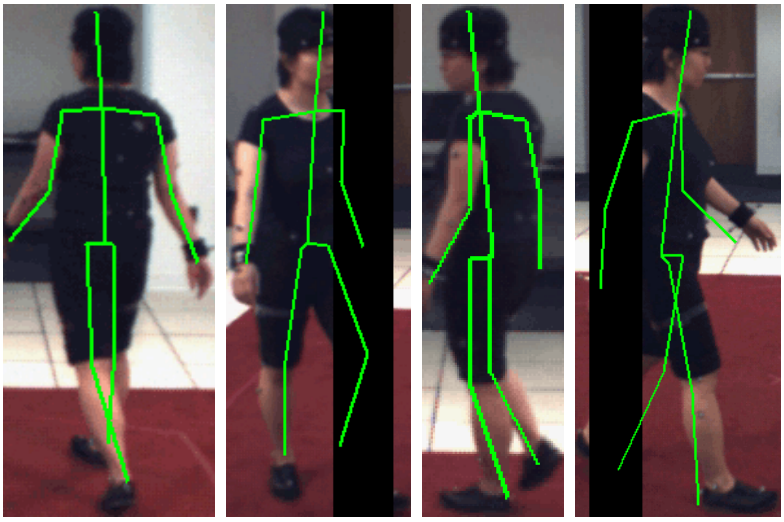
For each occlusion condition, the table shows the increase in error over all evaluated actions and subjects. For the vertical conditions, each of which occludes 20% of the image, the average error is approximately 7% higher. The horizontal settings result in a 9% increase, while occlusion covers one third of the observation. There are, however, large differences in performance for different trials. We will discuss these in the next section.

For the *pole* sequence, we obtained a 3D error of 43.54 mm averaged over all joints and relative to the pelvis joint. We also discuss these results in the next section.

4.5 Discussion

Several sequences show slightly lower errors when occlusion is added. This is most likely caused by bad foreground segmentation. In the occluded conditions, these distracting regions are ignored.

In general, we observe differences in accuracy between occlusion settings. The horizontal settings have a slightly higher error (9%) compared to the vertical occlusion settings (7%). This effect can be partly explained by the higher percentage



(a) (b) (c) (d)

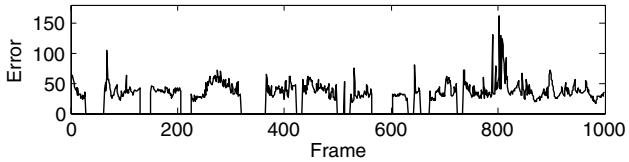
Fig. 5 Recovered poses of frames 250, 500, 750, and 1000. In frames 500 and 1000, the pole is shown, but a larger region is not taken into account (60 and 40%, respectively). The 3D pelvis location was set manually, as we only estimate relative joint locations.

of occlusion (33%, compared to 20%). However, there are differences between settings. Most of these differences are caused by several individual trials. We will discuss these sequences that largely contribute to the increased error.

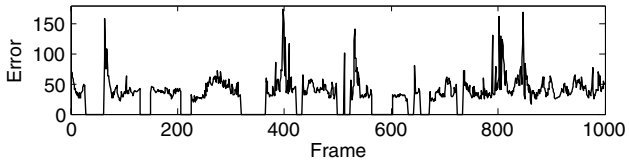
A significant part of the increase in recovery error is due to the boxing action. Especially for subject 3, almost all occlusion settings result in significantly higher errors. Analysis of these results reveal that many examples from the gesture action are selected. For the gesture action, a similar trend is visible, especially in the *h_bottom* condition. We expect that this is mainly due to the fact that subject 3 wears dark clothes, which results in relatively few edge responses between arms and body. Therefore, examples from beckoning gestures and box punches are selected interchangeably. The same effect is visible for subject 2, but to a much lesser extent.

For the jog action, recovery accuracy is significantly lower in the *h_bottom* occlusion setting. As the hands do not move a lot while jogging, the legs are most informative in this case as well.

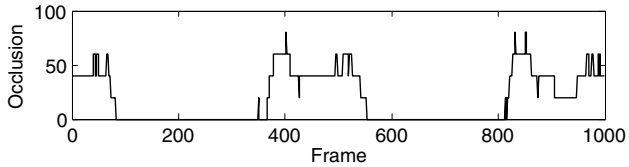
The throwing and catching trials of subject 2 have a relatively low increase in error, whereas the *h_bottom* condition for subject 3 shows much higher errors. Similar to earlier observations, the relatively low number of edge responses for subject 3 is probably the reason that the missing edges for the feet result in decreased accuracy. Also, the *v_left* occlusion setting shows higher error values, which can be explained since the subjects both throws and catches the ball right to him. In the image, this corresponds to the left side of the observation.



(a) No occlusion



(b) Occlusion simulated (*pole* sequence)



(c) Percentage of occlusion

Fig. 6 Mean relative 3D error (in mm) plots for HumanEva-I Walking, performed by subject 1 and viewed with a single camera (C1), (a) without occlusion, (b) for *pole* condition, and (c) percentage of occlusion for *pole* condition. Instances that have a zero error contain invalid mocap.

The error plots for the *pole* sequence are shown in Figure 6. The amount of occlusion for each frame is in the range of 0–80%. On average, 19.06% of the observation is occluded. The graphs clearly show that the error increases under occlusion. The average increase in error is 11.56%. It should be noted that this is mainly due to those frames where more than half of the observation is occluded. Here, the average error is 57.00 mm, compared to 39.17 mm for the non-occluded sequence. When occlusion is in the range of 20–40%, the mean error is 43.63 mm compared to 37.49 mm for the original. Differences between the upper two graphs for those frames where no occlusion is present are due to differences in normalization. The pose estimates for several frames are shown in Figure 5.

4.5.1 Comparison with Related Discriminative Work

The HumanEva dataset has been used by others to evaluate their work. In Table 3, we compare previous reports on the HumanEva-I dataset, obtained using discriminative approaches. Only non-occluded observations are used. The errors in Table 3 are indicative, as there are various differences between methods. First, Elgammal and Lee [15, 34] normalize the errors for rotation. Second, some works evaluate their approach on the validation sets, which allows for parameter tuning since the online evaluation system is not required. Also, the observations might be closer to

Table 3 Comparison of results of discriminative approaches, reported on HumanEva-I. Dynamics (*dyn.*) indicates whether a dynamical model (possibly activity-specific) is employed. All errors are relative to the *torsoDistal* joint. Errors in 3D are in *mm*, for 2D in pixels. Direct comparison is hindered due to differences between evaluations (different subjects, validation set instead of test set, only part of the sequence). [15, 34] use a rotation-normalized error measure.

	Image representation	Dyn.	Walking	Box	2D/3D
Bo et al. [6]	Histogram of SC	N	25.66	30.40	3D
Bo and Sminchisescu [5]	HOG descriptor	N	37.07	89.47	3D
Elgammal and Lee [15]	Silhouette	Y	31.36		3D
Lee and Elgammal [34]	Silhouette	N	76.56		3D
Okada and Soatto [50]	HOG descriptor	N	37.98		3D
Poppe (this work)	HOG descriptor	N	45.05	94.97	3D
Poppe [56]	HOG descriptor	N	45.36	94.48	3D
Urtasun et al. [78]	Hierarchical features	N	32.70	38.50	3D
Howe [29]	Silhouette	Y	15.00		2D
Urtasun et al. [78]	Hierarchical features	N	5.18	6.68	2D

the training set as training and validation sets are obtained from the same movement sequence. Third, different sets are used for training. Several authors have used person-specific training sets (e.g. [78]). Such an approach does not give any information about the ability to generalize over subjects.

Howe [29] presents an example-based approach and uses a batch approach to recover the most likely poses over a sequence of observations. This allows to avoid forward–backward ambiguities, which are common when using silhouettes. Our approach is closely related to Poppe's [56], who uses a global normalization and matching distance. Differences in accuracy between a global and cell-level normalization are small. Moreover, our approach can deal with partial occlusions.

Regression-based approaches (e.g. [6, 50, 78]) in general show lower errors. However, this comes at the cost of a computationally costly training phase, which has proved difficult to generalize to larger numbers of training samples. Also, it remains to be investigated how these approaches can cope with partial occlusions.

5 Conclusion

We presented an approach to recover human poses directly from image observations, even under significant occlusions. Our method requires segmentation of the human subject in the image and prediction of the occluded areas. Recent work in the domain of human detection can produce this information. We take an example-based approach, where we encode the image observation using HOG. For each grid cell, we assign a weight that indicates whether the cell is occluded. As such, we can use the same example database, regardless of the type of occlusion. Experiments were performed on the HumanEva-I dataset, which was also used to construct a database of image–pose pairs. Since the dataset does not contain occlusions, we

additionally simulated occlusion for six different conditions. Poses are estimated by interpolating the poses corresponding to the 25 closest matches from a database. In this matching, we treat the occluded cells as missing observations. This allows us to recover poses even if a significant percentage of the image observation is occluded.

For non-occluded observations, we report 3D errors of approximately 69 mm, relative to the pelvis and averaged over a set of 20 joint locations. For walking and jogging, this error is approximately 45 mm. Our results show approximately 10% increase in error when 20% of the observation is occluded. When 33% of the observation is occluded, the error is on average 15% higher compared to the observations without occlusions. We plan to combine our work with human detection. One avenue of future research is to analyze how stereo information can aid in the human detection and occlusion prediction tasks. Also, we are looking at ways to reduce the linear complexity in the number of examples, either using hashing or a regression-based approach.

Acknowledgment. We wish to thank the authors of [69] for making their database available.

References

1. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3851, pp. 50–59. Springer, Heidelberg (2006)
2. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 28(1), 44–58 (2006)
3. van den Bergh, M., Koller-Meier, E., van Gool, L.J.: Real-time body pose recognition using 2D or 3D haarlets. International Journal of Computer Vision (IJCV) 83(1), 72–84 (2009)
4. Bissacco, A., Yang, M.-H., Soatto, S.: Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN, June 2007, pp. 1–8 (2007)
5. Bo, L., Sminchisescu, C.: Twin Gaussian processes for structured prediction. International Journal of Computer Vision (IJCV) (2009) (to appear)
6. Bo, L., Sminchisescu, C., Kanaujia, A., Metaxas, D.: Fast algorithms for large scale conditional 3D prediction. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
7. Bosch, A., Zisserman, A., Muñoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the International Conference on Image and Video Retrieval (CIVR 2007), Amsterdam, The Netherlands, July 2007, pp. 401–408 (2007)
8. Bowden, R., Mitchell, T.A., Sarhadi, M.: Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences. Image and Vision Computing 18(9), 729–737 (2000)
9. Brand, M.: Shadow puppetry. In: Proceedings of the International Conference on Computer Vision (ICCV 1999), Kerkyra, Greece, September 1999, vol. 2, pp. 1237–1244 (1999)

10. Chakraborty, B., Rudovic, O., Gonzàlez, J.: View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR 2008), Amsterdam, The Netherlands, September 2008, pp. 1–6 (2008)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, June 2005, vol. 1, pp. 886–893 (2005)
12. Dong, L., Parameswaran, V., Ramesh, V., Zoghlami, I.: Fast crowd segmentation using shape indexing. In: Proceedings of the International Conference On Computer Vision (ICCV 2007), Rio de Janeiro, Brazil, October 2007, pp. 1–8 (2007)
13. Ek, C.H., Rihan, J., Torr, P.H.S., Rogez, G., Lawrence, N.D.: Ambiguity modeling in latent spaces. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 62–73. Springer, Heidelberg (2008)
14. Elgammal, A.M., Davis, L.S.: Probabilistic framework for segmenting people under occlusion. In: Proceedings of the International Conference On Computer Vision (ICCV 2001), Vancouver, Canada, July 2001, vol. 2, pp. 145–152 (2001)
15. Elgammal, A.M., Lee, C.-S.: Trackingpeople on a torus. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31(3), 520–538 (2009)
16. Fathi, A., Mori, G.: Human pose estimation using motion exemplars. In: Proceedings of the International Conference On Computer Vision (ICCV 2007), Rio de Janeiro, Brazil, Ocotber 2007, pp. 1–8 (2007)
17. Pedro, F.: Felzenszwalb, David McAllester, and Deva Ramanan. Discriminatively trained multiscale deformable part models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
18. Fossati, A., Arnaud, E., Horaud, R., Fua, P.: Tracking articulated bodies using generalized expectation maximization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2008), Washington, DC, June 2008, pp. 1–6 (2008)
19. William, T.: Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In: Proceedings of the Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, June 1995, pp. 296–301 (1995)
20. Gandhi, T., Trivedi, M.M.: Image based estimation of pedestrian orientation for improving path prediction. In: Proceedings of the Intelligent Vehicles Symposium (IV 2008), Eindhoven, The Netherlands, June 2008, pp. 506–511 (2008)
21. Gavrila, D.M.: A Bayesian, exemplar-based approach to hierarchical shape matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29(8), 1408–1421 (2007)
22. Gond, L., Sayd, P., Chateau, T., Dhome, M.: A 3D shape descriptor for human pose recovery. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2008. LNCS, vol. 5098, pp. 370–379. Springer, Heidelberg (2008)
23. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3D structure with a statistical image-based shape model. In: Proceedings of the International Conference on Computer Vision (ICCV 2003), Nice, France, October 2003, vol. 1, pp. 641–647 (2003)
24. Guo, F., Qian, G.: Human pose inference from stereo cameras. In: Proceedings of the Workshop on Applications of Computer Vision (WACV 2007), Austin, TX, February 2007, p. 37 (2007)

25. Gupta, A., Chen, T., Chen, F., Kimber, D., Davis, L.S.: Context and observation driven latent variable model for human pose estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
26. Howe, N.R.: Silhouette lookup for automatic pose tracking. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2004), Los Alamitos, CA, June 2004, p. 15 (2004)
27. Howe, N.R.: Flow lookup and biological motion perception. In: Proceedings of the Internation Conference on Image Processing (ICIP 2005), Genova, Italy, September 2005, vol. 3, pp. 1168–1171 (2005)
28. Howe, N.R.: Boundary fragment matching and articulated pose under occlusion. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, pp. 271–280. Springer, Heidelberg (2006)
29. Nicholas, R.: Howe. Recognition-based motion capture and the HumanEva II test data. In: Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation, Minneapolis, MN (June 2007)
30. Isard, M., Blake, A.: CONDENSATION - conditional density propagation for visual tracking. International Journal of Computer Vision 29(1), 5–28 (1998)
31. Kanaujia, A., Sminchisescu, C., Metaxas, D.: Semi-supervised hierarchical models for 3D human pose reconstruction. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN, June 2007, pp. 1–8 (2007)
32. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, June 2004, vol. 2, pp. 506–513 (2004)
33. Stephen, J.: Krotosky and Mohan M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 8(4), 619–629 (2007)
34. Lee, C.-S., Elgammal, A.M.: Simultaneous inference of view and body pose using torus manifolds. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2006), Kowloon Tong, Hong Kong, August 2006, vol. 3, pp. 489–494 (2006)
35. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, June 2004, vol. 2, pp. 53–60 (2004)
36. Lin, Z., Davis, L.S.: A pose-invariant descriptor for human detection and segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 423–436. Springer, Heidelberg (2008)
37. Lin, Z., Davis, L.S., Doermann, D., DeMenthon, D.: Hierarchical part-template matching for human detection and segmentation. In: Proceedings of the International Conference on Computer Vision (ICCV 2007), Rio de Janeiro, Brazil, October 2007, pp. 1–8 (2007)
38. Liu, X., Yu, T., Sebastian, T., Tu, P.: Boosted deformable model for human body alignment. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60(2), 91–110 (2004)
40. Lu, W.-L., Little, J.J.: Simultaneous tracking and action recognition using the PCA-HOG descriptor. In: Proceedings of the Canadian Conference on Computer and Robot Vision (CRV 2006), Quebec City, Canada, June 2006, p. 6 (2006)

41. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision (IJCV) 60(1), 63–86 (2004)
42. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(10), 1615–1630 (2005)
43. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
44. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23(4), 349–361 (2001)
45. Mori, G., Malik, J.: Recovering 3D human body configurations using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 28(7), 1052–1062 (2006)
46. Navaratnam, R., Fitzgibbon, A.W., Cipolla, R.: Semi-supervised learning of joint density models for human pose estimation. In: Proceedings of the British Machine Vision Conference (BMVC 2006), Edinburgh, United Kingdom, September 2006, vol. 2, pp. 679–688 (2006)
47. Niebles, J.C., Han, B., Ferencz, A., Fei-Fei, L.: Extracting moving people from internet videos. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 527–540. Springer, Heidelberg (2008)
48. Ning, H., Hu, Y., Huang, T.S.: Efficient initialization of mixtures of experts for human pose estimation. In: Proceedings of the International Conference on Image Processing (ICIP 2008), San Diego, CA, October 2008, pp. 2164–2167 (2008)
49. Ning, H., Xu, W., Gong, Y., Huang, T.S.: Discriminative learning of visual words for 3D human pose estimation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
50. Okada, R., Soatto, S.: Relevant feature selection for human pose estimation and localization in cluttered images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 434–445. Springer, Heidelberg (2008)
51. Ong, E.-J., Gong, S.: A dynamic 3D human model using hybrid 2D-3D representations in hierarchical pca space. In: Proceedings of the British Machine Vision Conference (BMVC 1999), Nottingham, United Kingdom, September 1999, pp. 33–42 (1999)
52. Ong, E.-J., Micilotta, A.S., Bowden, R., Hilton, A.: Viewpoint invariant exemplar-based 3D human tracking. Computer Vision and Image Understanding (CVIU) 104(2-3), 178–189 (2006)
53. Onishi, K., Takiguchi, T., Ariki, Y.: 3D human posture estimation using the HOG features from monocular image. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2008), Tampa, FL, December 2008, pp. 1–4 (2008)
54. Peursum, P., Venkatesh, S., West, G.: Observation-switching linear dynamic systems for tracking humans through unexpected partial occlusions by scene objects. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2006), Kowloon Tong, Hong Kong, August 2006, vol. 4, pp. 929–934 (2006)
55. Peursum, P., Venkatesh, S., West, G.: Tracking-as-recognition for articulated full-body human motion analysis. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN, June 2007, pp. 1–8 (2007)
56. Poppe, R.: Evaluating example-based pose estimation: Experiments on the HumanEva sets. In: Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (CVPR-EHuM), Minneapolis, MN (June 2007)

57. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding (CVIU) 108(1-2), 4–18 (2007)
58. Poppe, R., Poel, M.: Comparison of silhouette shape descriptors for example-based human pose recovery. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR 2006), Southampton, United Kingdom, April 2006, pp. 541–546 (2006)
59. Poppe, R., Poel, M.: Body-part templates for recovery of 2D human poses under occlusion. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2008. LNCS, vol. 5098, pp. 289–298. Springer, Heidelberg (2008)
60. Porikli, F.: Integral histogram: A fast way to extract histograms in cartesian spaces. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, June 2005, vol. 1, pp. 829–836 (2005)
61. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking people by learning their appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29(1), 65–81 (2007)
62. Rogez, G., Orrite-Uruñuela, C., del Rincón, J.M.: A spatio-temporal 2D-models framework for human pose recovery in monocular sequences. Pattern Recognition 41(9), 2926–2944 (2008)
63. Rogez, G., Rihan, J., Ramalingam, S., Orrite-Uruñuela, C., Torr, P.H.S.: Randomized trees for human pose detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
64. Rosales, R.E., Sclaroff, S.: Learning body pose via specialized maps. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, December 2001, vol. 14, pp. 1263–1270 (2001)
65. Rosales, R.E., Sclaroff, S.: Combining generative and discriminative models in a framework for articulated pose estimation. International Journal of Computer Vision (IJCV) 67(3), 251–276 (2006)
66. Rosales, R.E., Siddiqui, M., Alon, J., Sclaroff, S.: Estimating 3D body pose using uncalibrated cameras. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, December 2001, vol. 1, pp. 821–827 (2001)
67. Shakhnarovich, G., Viola, P.A., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: Proceedings of the International Conference on Computer Vision (ICCV 2003), Nice, France, October 2003, vol. 2, pp. 750–759 (2003)
68. Sigal, L., Bălan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, December 2008, vol. 20, pp. 1337–1344 (2008)
69. Sigal, L., Black, M.J.: HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI (September 2006)
70. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Learning joint top-down and bottom-up processes for 3D visual inference. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, June 2006, vol. 2, pp. 1743–1752 (2006)
71. Sminchisescu, C., Kanaujia, A., Metaxas, D.N.: BM^3E: Discriminative density propagation for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29(11), 2030–2044 (2007)
72. Sullivan, J., Carlsson, S.: Recognizing and tracking human action. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 629–644. Springer, Heidelberg (2002)

73. Sun, Y., Bray, M., Thayananthan, A., Yuan, B., Torr, P.H.S.: Regression-based human motion capture from voxel data. In: Proceedings of the British Machine Vision Conference (BMVC 2006), Edinburgh, United Kingdom, September 2006, vol. 1, pp. 277–286 (2006)
74. Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P.H.S., Cipolla, R.: Pose estimation and tracking using multivariate regression. Pattern Recognition Letters 29(9), 1302–1310 (2003)
75. Thurau, C.: Behavior histograms for action recognition and human detection. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 271–284. Springer, Heidelberg (2007)
76. Toyama, K., Blake, A.: Probabilistic tracking with exemplars in a metric space. International Journal of Computer Vision 48(1), 9–19 (2002)
77. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2008)
78. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity-independent human pose inference. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, June 2008, pp. 1–8 (2008)
79. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, December 2001, vol. 1, pp. 511–518 (2001)
80. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. International Journal of Computer Vision (IJCV) 75(2), 247–266 (2007)
81. Wu, B., Nevatia, R.: Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. International Journal of Computer Vision (IJCV) 82(2), 185–204 (2009)
82. Xu, L.-Q., Hogg, D.C.: Neural networks in human motion tracking - an experimental study. Image and Vision Computing 15(8), 607–615 (1997)
83. Yang, H.-D., Lee, S.-W.: Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. Pattern Recognition 40(11), 3120–3131 (2007)
84. Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T.: Fast human detection using a cascade of histograms of oriented gradients. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, June 2006, vol. 2, pp. 1491–1498 (2006)

Part IV
Architectures for Distributed Agent-Actor Communities