Full Body Pose Estimation During Occlusion using Multiple Cameras

Fihl, Preben; Cosar, Serhan

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

CVMT Technical Report; 10-02 ISSN/ISBN: 1601-3646

Full Body Pose Estimation During Occlusion using Multiple Cameras

P. Fihl and S. Cosar

Computer Vision and Media Technology Lab.
Aalborg University, Denmark
fihl@create.aau.dk

Faculty of Engineering and Natural Sciences
Sabanci University, Tuzla, Istanbul, Turkey
serhancosar@su.sabanciuniv.edu

Abstract. Automatic estimation of the human pose enables many interesting applications and has therefore achieved much attention in recent years. One of the most successful approaches for estimating unconstrained poses has been the pictorial structures framework. However, occlusions between interacting people is a very challenging problem for methods based on pictorials structure as for any other monocular pose estimation method. In this report we present work on a multi-view approach based on pictorial structures that integrate low level information from multiple calibrated cameras to improve the 2D pose estimates in each view. The proposed method in shown to work under heavy occlusions but does not improve the pose estimates in the non-occluded cases in it's current form.

1 Introduction

Automatic estimation of the human pose enables many interesting applications and has therefore achieved much attention in recent years. Accurate pose estimations can give a good description of the actions being performed in a video and hence be used for *e.g.*, automatic video surveillance, human-computer interaction or automatic video annotation. In this report we present a method to do full body pose estimation by combining information from multiple cameras to deal with the problem of occlusions between people.

One of the most successful approaches for estimating unconstrained poses has been the pictorial structures framework [7] which has been improved and extended in a number of works [2, 8, 18]. One of the main challenges for human pose estimation is the inherent problem of occlusions, especially self occlusions, i.e. one body part occluding another body part of the same person. Several methods have been proposed to deal with this problem by modeling the self occlusions directly in the body model [19, 20] or by utilizing a foreground mask and maximizing the area of foreground covered by body parts [15].

The problem of occlusions increases when multiple people interact. [6] proposes a multi-person pictorial structures model and estimate the front-to-back ordering of people to find the probability of occlusions between people. [9] also addresses the problem of occlusions by other foreground objects and propose to detect occluded body parts by pruning the foreground mask into a mask of possible occlusions and then include these detections in the inference process of the pictorial structures.

All of these approaches use a single viewpoint to provide a 2D pose estimate but severe occlusions will always cause problems for monocular methods in general. Combining information from multiple cameras can help to solve this issue and a number of multi-view methods for reconstruction of the human pose have been proposed in recent years, based on for example visual hull [4, 12] or skeleton models [3, 5].

The multi-view methods range from carefully calibrated studio setups to surveillance setups with the placement of the cameras determined by the environment. The studio setups usually deal with human pose estimation in 3D and the cameras are arranged for that purpose specifically. [3–5, 12, 17] are examples of such methods and [11] presents a representative multi-view data set for human pose estimation and action recognition with references to related methods and data sets. The surveillance-like setups utilize multiple overlapping views in unconstrained environments and these methods typically deal with the problem of tracking people through occlusions rather than pose estimation, like in [10, 16, 21]. The video data in surveillance setups does rarely cover the scenes as thoroughly as studio setups and the video is not specifically captured with pose estimation as a goal making it a difficult task. However, in this report we present a method to combine information from multiple overlapping cameras in a surveillance-like setup into 2D full body pose estimates thereby effectively dealing with the problem of occlusions in single-view approaches.

Other approaches address a similar problem. [13] do full body pose estimation by combining the body part likelihoods from two cameras in an iterative approach. [14] use silhouette based shape matching in each view and then project the best mathing shape models to the other views where a multi-view matching score is calculated. Our method also integrates low level information, like the body part likelihoods, but we build on the pictorial structures framework which has shown good results in single-view approaches and we do the integration in a non-iterative way. The use of the pictorials structures framework also means that we are not limited to a set of poses represented in a database as in [14].

Our multi-view pose estimation process will not attempt to generate a 3D pose estimate. Instead, our method will integrate information from multiple views to get improved 2D pose estimations. There are no restrictions on the camera setup which fits well with the typical surveillance scenario. The method will integrate information from a variable number of cameras and does not require a person or a body part to be visible in all cameras.

The rest of the report is organized as follows. First, section 2 gives a brief description of the single-view pictorial structures framework. Section 3 then describes the integration of pose estimation from multiple cameras. Section 4 presents results using the proposed method which are discussed in section 5. Finally, section 6 concludes the report.

2 Single-view pictorial structures framework

Our approach extends the single-view pose estimation method presented in [9] which build on the pictorial structures framework of [18]. The single-view pose estimation method in [9] will be summarized next.

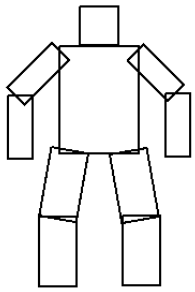

Fig. 1. Pictorial structure representation that models the human body.

The pose estimation is based on detections of individual body parts which are combined into body configurations under kinematic constraints in a pictorial structures framework. The body model contains ten body parts but only six types of body parts are detected since left and right limbs are assumed to have the same appearance and shape. The body part types are heads, torsos, upper arms, lower arms, upper legs, and lower legs (see figure 1).

A robust background subtraction is first applied to limit the search space for body parts to foreground regions. The body parts are detected using both person-specific appearance models and generic edge-based shape models. The two types of information are combined with a dynamic weighting dependant on the local quality of the appearance information.

The initialization of the appearance models is done by detecting isolated people in a characteristic walking pose. The detection of body parts in the initialization process is done only with edge-based detectors but constraining the initialization phase to one characteristic pose ensures good detections despite the relatively weak detectors. Initialization can alternatively be done by clustering edge-based body part detections [18]. This requires a set of frames for initialization but does not require people to be in the characteristic walking pose.

A set of pairwise kinematic constraints are applied to the individual body parts detections. These are represented as a tree-structured model which allow for efficient calculation of the posterior probabilities. At this stage the number of visible body parts is unknown and a set of body configurations with one arm and one leg is drawn from the posterior of the pictorial structure. The true modes of the posterior (representing the poses of visible body parts) are then found using a mean shift approach to produce the final pose estimate.

3 Multi-view pose integration

The body part detection of the single-view pose estimation method produces a probability map for each type of body part. By combining these probability maps from multiple cameras we will get a fusion of low level data that will ideally improve the results of the inference in the tree-structured body model and the subsequent estimation of the

final pose, especially when body part detections from one view are poor as a result of occlusion. Figure 2 shows how our approach fits into the single-view approach.

To enable the integration of the body part probability maps we use calibrated cameras. This allows us to calculate the projections of probability maps from different cameras into a common world coordinate system (illustrated in figure 3). By sampling new body parts from this 3D space and projecting them back to each view we get new enhanced 2D body part probability maps. In essence, our multi-view pose estimation method combines body part probability maps from multiple cameras and generates improved probability maps for each camera. The rest of the pose estimation proceeds as in the single-view method.

The body part probability maps are confined by the foreground masks which ensures that the overlap between their projections in world coordinates is within a small volume. However, when multiple people interact the foreground always contain more than one person, so the foreground mask merely defines a region of interest and not the silhouette of each person.

The probability maps express the probability of a body part given its 2D orientation. When projecting these into 3D we would ideally want to know the third rotation angle (out of the image plane) to ensure that the probabilities are only combined with the corresponding 3D orientation from the other views. This is however not possible.One may try to combine all possible 3D orientations for a given 2D orientation with all 2D orientations in the other views. Rather than such a complex approach, we change representation to joints instead of body parts, *i.e.*, we transform each body part probability map into two joint probability maps. For instance, upper arm probability map is transformed into shoulder and elbow probability maps. By describing a body part by its two end points we get the locations of the corresponding joints. Joint probability maps can now be generated by letting the probability of both joints be the same as the probability of the corresponding body part.

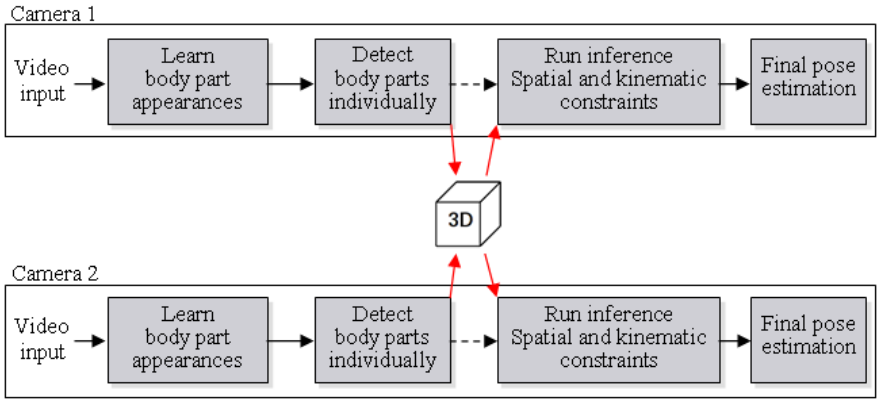

Fig. 2. A block diagram that shows how our approach fits into the single-view approach for combining information from two cameras.

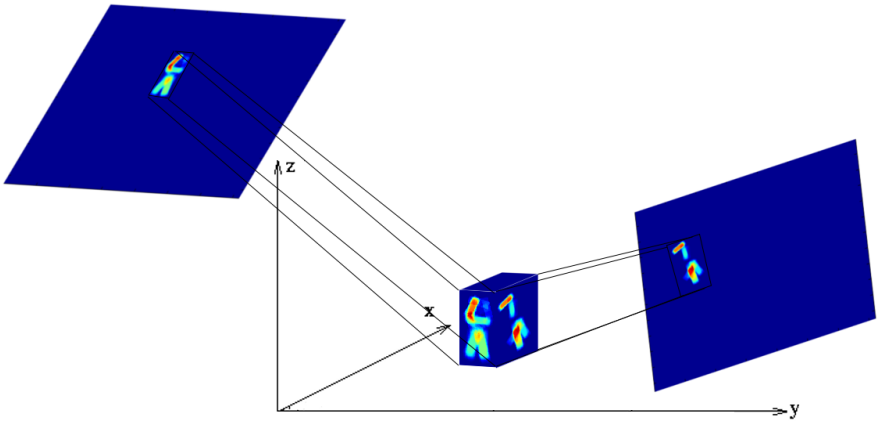

Fig. 3. Probability maps from different viewpoints are combined by projecting them into a common world space (the illustration is not a correct perspective projection). The maps illustrate the probabilities of arms and legs with dark blue corresponding to zero probability and red corresponding to high probability.

The joint probabilities $P(J)$ are independent of orientation (θ) so we take the maximum of body part probabilities $P(Bp)$ over all orientations:

$$P(J|x,y) = max(P(Bp|x',y',\theta)), \forall\theta \tag{1}$$

where x and y are image coordinates of the joint and x' and y' are the image coordinates of the corresponding body part center.

With known camera calibrations we can now project the non-zero joint probabilities into the 3D world space where it will represent a volume. This volume can be combined with the projections of probabilities of the same joint from other cameras. The probabilities are added together where ever the volumes of the projections overlap.

The projection of joint probabilities corresponds to a cylindrical volumes in 3D. In world space however, we want joints to be represented as spheres so we apply mean shift clustering to divide the combined volume into a set of clusters, each cluster representing a joint location, and the probability of that joint being the maximum probability of points belonging to the cluster.

This process is done for each joint individually. The goal is however to generate improved *body part* probability maps for each view, so at this stage we will connect top and bottom joints to get a body part representation, for example, we connect a shoulder joint (top joint) to an elbow joint (bottom joint) to get an upper arm sample. By back-projecting many body part samples, we can approximate the corresponding probability maps in each view.

Two joint-clusters are connected if the distance between them corresponds to the length of the appropriate body part $\pm 20\%$ of the length. Body part lengths are estimated for each person in the initialization phase. The estimation is based on a set of standard body part proportions that are scaled with the height of the person in the initialization pose. The joint-cluster are furthermore only connected if the area between the cluster

centers project back onto the foreground mask in all views. The second condition is necessary to minimize the risk of connecting joints that do not correspond to real body parts.

One sample is drawn from each connected cluster-pair and the probability of that body part sample is the sum of the probabilities of the two joints. All these 3D body part samples are projected back onto each view where they will be connected into body configurations through inference in the tree-structured body model as in the single-view approach. The final pose estimate is also found like in the single-view case, *i.e*. using a mean shift approach to find the true modes of the posterior.

4 Results

This section presents the results of the preliminary tests of the proposed integration of multiple views for pose estimation.

We test the method on a video sequence from the PETS 2009 data set [1]. The data set provides video from eight calibrated cameras with people interacting in the overlapping field of view of all cameras. We generate the pose estimation results from two cameras using the single-view approach [9] and compare these results to the pose estimates found by using the proposed integration of body part probability maps from multiple views. Both sets of results use the same body part detection procedure. The multi-view approach then integrates body part probability maps but then proceeds as in the single-view approach. The comparison is based on a qualitative analysis of example frames.

The single-view approach generally performs better in the non-occluded cases (figure 4 rows *a* and *b*). This will be discussed in the next section. The multi-view approach performs better in the occluded cases (figure 4 rows *c* and *d*), especially under full occlusion (figure 4 column 3).

The main effect of the proposed multi-view pose integration at this stage is the ability to transfer a good pose estimate from one view into an improved pose estimate in another view where the person is heavily occluded.

5 Discussion

A very important difference between the single-view approach and the presented multi-view approach is the assumptions about body part proportions. The single-view approach assume fixed body part proportions in the image plane. The scale of a body part is estimated during initialization, but after that the ratio of length to width is fixed. This means that the single-view approach at this stage does not handle foreshortening at all. The flexibility of the joints in the pictorial structures model and the mean shift approach for finding the true modes of the posterior allow the single-view approach to handle some foreshortening but not very much. When estimating the poses of walking people this rarely becomes a problem but if a person were to for example point at something close to the camera the foreshortening would cause the single-view approach to fail.

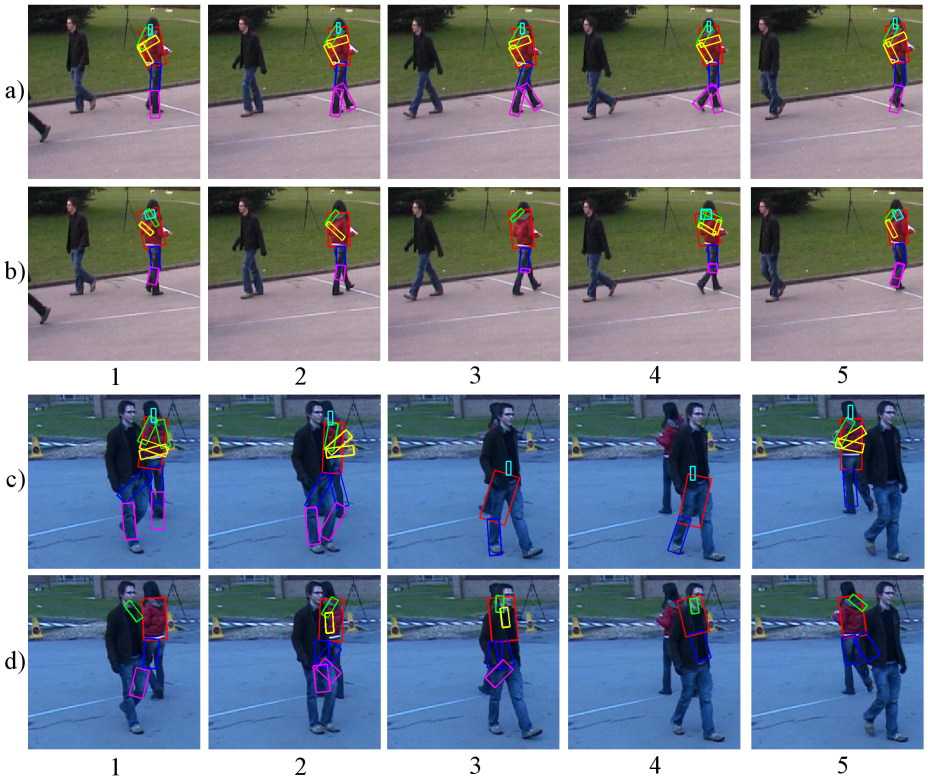

Fig. 4. Comparison of the multi-view approach against the single-view approach. Rows a) and c) show the results of pose estimation with the single-view approach. Rows b) and d) show the results of pose estimation with the multi-view approach. Each column of the figure shows a frame from two synchronized cameras.

The multi-view approach also assumes fixed body part proportions but in the 3D world space and not in the image plane. This means that the multi-view approach can handle the extreme foreshortening of a person pointing towards the camera as long as another camera is capturing the arm without the foreshortening. By keeping the constrain on body part proportions from the single-view approach would limit the multi-view approach from an interesting capability. The results show however that lifting the constrain of body part proportions in the image plane significantly increases the noise in the body part probability maps that are generated from back projections of the 3D body parts. The connection of 3D joints into body parts does not necessarily result in a body part corresponding to a sample from one of the views (this is what allow the handling of extreme foreshortening). However, many joint-pairs correspond to a foreshortened body part in one view also when there is no real foreshortening resulting in increased noise in the body part probability maps. This side effect significantly reduces the performance of the multi-view pose integration, an effect that is seen clearly in the non-occluded cases (figure 4 rows *a* and *b*).

When people undergo heavy occlusion the single-view pose estimation tend to collapse whereas the multi-view approach maintain a reasonable pose estimate. By combining the pose estimation with a person tracker it would be possible to predict when occlusions could occur and then rely on the single-view pose estimation in the non-occluded cases and include the multi-view pose integration when occlusions occur.

6 Conclusion

This report presents a multi-view extension to the pose estimation method of [9] based on the pictorial structures framework. Low level information about body part probabilities from multiple cameras are combined in 3D world space by utilizing calibrated cameras. The combined probabilities are projected back into each view to generate improved 2D pose estimates. Preliminary results indicates that the proposed method improves pose estimates under heavy occlusion compared to the single-view approach but it performs worse when there is no occlusion.

A number of alternatives to individual steps in the method could be explored further. The most interesting one would be to build the tree-structured body model in 3D rather than in 2D and then do the whole pose estimation in 3D world space. This would allow us to make much better use of the kinematic constraints and it would also result in one joint pose estimate that can be back projected to each view instead of multiple (possibly different) pose estimates in the different views. In terms of testing it would be very interesting to see how the method performs when integrating more than two views.

References

1. PETS 2009 Benchmark Data by University of Reading. http://www.cvg.rdg.ac.uk/PETS2009/a.html.
2. M. Andriluka, S. Roth, and B. Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In *CVPR*, 2009.
3. Daniel Chen, Pi-Chi Chou, Clinton B. Fookes, and Sridha Sridharan. Multi-view Human Pose Estimation using Modified Five-point Skeleton Model. In *Int. Conference on Signal Processing and Communication Systems*, Dec. 2007.
4. Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas Andriacchi. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. *International Journal of Computer Vision*, 87, 2010.
5. E de Aguiar, C. Theobalt, M. Magnor, and H-P. Seidel. Reconstructing Human Shape and Motion from Multi-view Video. In *European Conference on Visual Media Production*, Nov. 2005.
6. M. Eichner and V. Ferrari. We Are Family: Joint Pose Estimation of Multiple Persons. In *ECCV*, September 2010.
7. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1), Jan 2005.
8. Vittorio Ferrari, Manuel J. Marn-Jimnez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.
9. P. Fihl and T.B. Moeslund. Pose Estimation of Interacting People using Pictorial Structures. In *Advanced Video and Signal Based Surveillance*, September 2010.

10. Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera People Tracking with a Probabilistic Occupancy Map. *PAMI*, 30(2), February 2008.
11. N. GkaleIis, Hansung Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3DPost Multi-View and 3D Human Action/Interaction Database. In *Conference for Visual Media Production*, Nov. 2009.
12. Laetitia Gond, Patrick Sayd, Thierry Chateau, and Michel Dhome. A 3D Shape Descriptor for Human Pose Recovery. In *Articulated Motion and Deformable Objects*, volume 5098 of *LNCS*. Springer Berlin / Heidelberg, 2008.
13. A. Gupta, A. Mittal, and L.S. Davis. Constraint Integration for Efficient Multiview Pose Estimation with Self-Occlusions. *PAMI*, 30(3), March 2008.
14. M. Hofmann and D.M. Gavrila. Multi-view 3D Human Pose Estimation Combining Single-frame Recovery, Temporal Integration and Model Adaptation. In *CVPR*, June 2009.
15. H. Jiang. Human Pose Estimation Using Consistent Max-Covering. In *ICCV*, 2009.
16. Yusuke Matsumoto, Toshikazu Wada, Shuichi Nishio, Takehiro Miyashita, and Norihiro Hagita. Scalable and Robust Multi-people Head Tracking by Combining Distributed Multiple Sensors. *Intelligent Service Robotics*, 3(1), 2010.
17. J. R. Mitchelson and A. Hilton. Simultaneous Pose Estimation of Multiple People using Multiple-View Cues with Hierarchical Sampling. In *British Machine Vision Conference*, 2003.
18. D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. *PAMI*, 29(1), Jan 2007.
19. Leonid Sigal and Michael J. Black. Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation. In *CVPR*, 2006.
20. Yang Wang and Greg Mori. Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation. In *ECCV*, 2008.
21. Jian Yao and Jean-Marc Odobez. Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios . In *ECCV workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications*, Oct. 2008.