Motion-Based View-Invariant Articulated Motion Detection and Pose Estimation Using Sparse Point Features

Shrinivas J. Pundlik and Stanley T. Birchfield

Clemson University, Clemson, SC USA
{spundli, stb}@clemson.edu

Abstract. We present an approach for articulated motion detection and pose estimation that uses only motion information. To estimate the pose and viewpoint we introduce a novel motion descriptor that computes the spatial relationships of motion vectors representing various parts of the person using the trajectories of a number of sparse points. A nearest neighbor search for the closest motion descriptor from the labeled training data of human walking poses in multiple views is performed. This observational probability is fed to a Hidden Markov Model defined over multiple poses and viewpoints to obtain temporally consistent pose estimates. Experimental results on various sequences of walking subjects with multiple viewpoints demonstrate the effectiveness of the approach. In particular, our purely motion-based approach is able to track people even when other visible cues are not available, such as in low-light situations.

1 Motivation for Articulated Human Motion Analysis

The detection of articulated human motion finds applications in a large number of areas such as pedestrian detection for surveillance, or traffic safety, gait/pose recognition for human computer interaction, videoconferencing, computer graphics, or for medical purposes. Johansson's pioneering work on moving light displays (MLDs) [1] has enabled researchers to study the mechanism and development of human visual system with a different perspective by decoupling the motion information from all other modalities of vision such as color and texture. One compelling conclusion that can be drawn from these studies is that motion alone captures a wealth of information about the scene. Others have made a similar observation [2, 3].

Figure 1 shows some examples of humans walking as seen from multiple angles along with their motion trajectories. Even though the appearance features (shape, color, texture) can be discriminative for detection of humans in the sequence, the motion vectors corresponding to the point features themselves can be used to detect them. The motion of these points becomes even more compelling when viewed in a video, as the human visual system fuses the information temporally to segment human motion from the rest of the scene. It is common knowledge that in spite of having a separate motion, each body part moves in a particular pattern. Our goal is to exploit the motion properties of the sparse points attached to a human body in a top-down approach for human motion analysis. More specifically, our attempt is to answer the question: If provided

only with the motion tracks (sparse point trajectories) and no appearance information, how well can an algorithm detect, track, and estimate the pose of a walking human in a video?

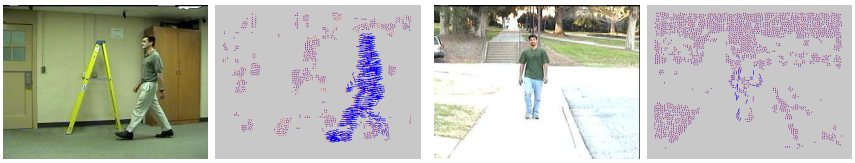

Fig. 1. Two examples of human walking motion at different viewing angles, and the motion vectors of the tracked feature points.

Previous work related to human motion detection and analysis can be loosely classified into three categories: pedestrian detection for surveillance, pose estimation, and action recognition. The nature of the algorithms dealing with the different categories varies significantly due to the differences in the input image sequences. Approaches for pedestrian detection are either appearance-based [4–6], use both appearance and stereo [7], or are based on modeling the periodic motion [8]. In contrast to pedestrian detection, human pose estimation [9–16, 3, 17–19] requires greater detail of the human motion to be captured, with a model that accounts for the disparate motions of the individual body parts. A related area of research is human action recognition [20, 21], in which the objective is to classify the detected human motion into one of several predefined categories using off-line training data for learning these action categories.

Even while considering only a single action category such as walking, human motion analysis remains a challenging problem due to various factors such as pose, scale, viewpoint, and scene illumination variations. Most approaches use appearance cues to perform human motion analysis, but these will not work when appearance information is lacking (e.g., at night in poorly lit areas). The few approaches that are predominantly motion based [3, 18] are limited in terms of viewpoint and lighting variations. In this paper, using only the sparse motion trajectories and a *single gait cycle* of 3D motion capture data points of a walking person for training, we demonstrate detection and pose estimation of articulated motion on various sequences that involve viewpoint, scale, and illumination variations, as well as camera motion. Our focus is on a top-down approach, where instead of learning the motion of individual joints and limbs as in [3], we learn the short-term motion pattern of the entire body in multiple pose and viewpoint configurations. Pose estimation can then be performed by a direct comparison of the learned motion patterns to those extracted from the candidate locations. The advantage of using such a top-down approach is that it greatly simplifies the learning step, facilitating one-shot learning. At the same time, the learned motion patterns can be reliably used to estimate the pose and the viewpoint in the presence of noise.

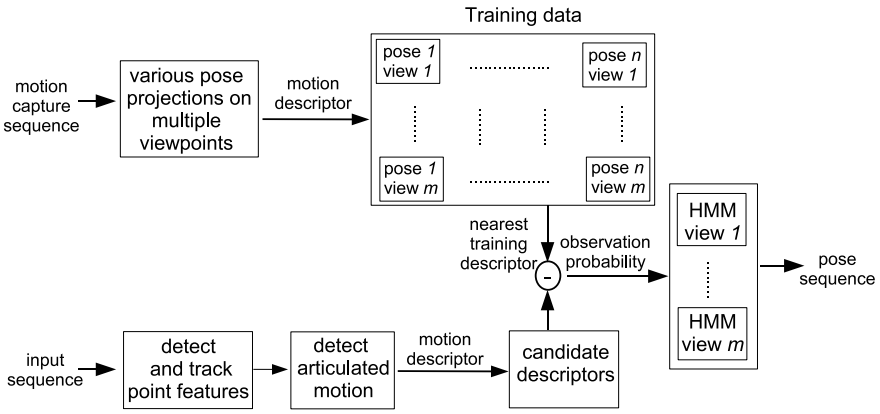

Fig. 2. Overview of the proposed approach to extract human motion models.

2 Learning Models for Multiple Poses and Viewpoints

An overview of the proposed approach is shown in Figure 2. Given an image sequence our goal is to segment, track, and determine the configuration of the walking human subject (2D pose and viewpoint) using only the sparse motion vectors corresponding to the feature points in the sequence. The primary reason for using sparse optical flow obtained from the tracked point features instead of a dense flow field for motion representation is efficiency of computation. The point features are detected and tracked using the Lucas-Kanade algorithm. Since there is a significant amount of self-occlusion, many point features representing the target are lost. Therefore, we use only short term feature trajectories between two consecutive frames. Let $V_t = \left(\mathbf{v}_1^{(t)}, \ldots, \mathbf{v}_k^{(t)}\right)$ be the tuple that describes the velocities of the k feature points at frame $t,\ t = 0, \ldots, T$, where $T + 2$ is the total number of frames in the sequence. The configuration of the subject in the current frame is denoted by $c_t = (m_t, n_t)$, where m_t and n_t are the 2D pose and view at time t, respectively. We assume that the viewpoint stays the same throughout the sequence. The configuration in the current frame is dependent not only on the motion vectors in the current frame but also on the configuration in the previous time instants. For determining c_t, the Bayesian formulation of the problem is given by

$$p(c_t|V_t, c_{0:t-1}) \propto p(V_t|c_{0:t})p(c_t|c_{0:t-1}), \tag{1}$$

where $p(V_t|c_{0:t})$ is the likelihood of observing the particular set of motion vectors given the configurations up to time t, and $p(c_t|c_{0:t-1})$ is the prior for time instant t that depends on previous configurations. Assuming a Markov process, we can write the above equation as

$$p(c_t|V_t, c_{0:t-1}) \propto p(V_t|c_t)p(c_t|c_{t-1}). \tag{2}$$

The estimate of the configuration at time t is $\hat{c}_t$, and our goal is to estimate configurations over the entire sequence, $\mathcal{C} = (\hat{c}_0, \ldots, \hat{c}_T)$. Learning the motion patterns of

the multiple poses and viewpoints involves obtaining a set of motion descriptors that describe each pose in each viewpoint first in the training data. The test data is then processed in a similar manner to obtain motion descriptors that are compared with the training data to obtain the likelihood of observing a particular pose and viewpoint configuration.

2.1 Training Data

For training, we used a single sequence from the CMU Motion Capture (mocap) data[1] in which the human subject is walking. A single gate cycle was extracted from the sequence. The obtained marker locations associated with the joints and limbs were projected onto simulated image planes oriented at various angles with respect to the subject for each pose (i.e., gait phase), and the corresponding motion vectors were obtained. A similar multi-view training approach was also adopted in [18]. The advantage of using the 3D data is that a single sequence provides a large amount of training data. Note that even though the motion capture data were obtained by calibrated cameras, our technique does not require any calibration since standard cameras have near unity aspect ratio, zero skew, and minimal lens distortion.

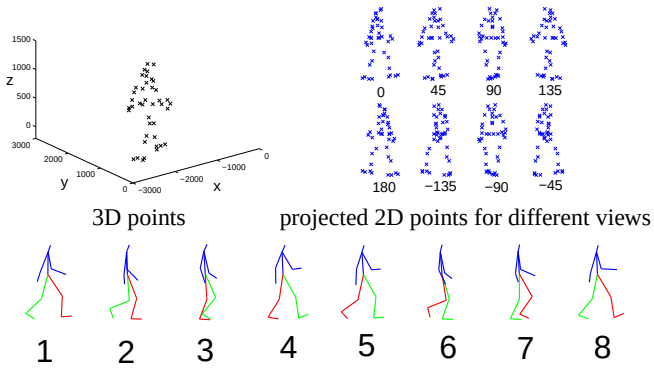

Fig. 3. Top: 3D Motion capture data and its projection onto various planes to provide multiple views in 2D. Bottom: Stick figure models for a sequence of poses (gait phases) for the profile view.

All possible views and poses are quantized to a finite number of configurations. Let M be the number of poses and N the number of views. Let $\mathbf{q}_m^{(i)} = (q_x^{(i)}, q_y^{(i)}, q_z^{(i)})^T$, be the 3D coordinates of the ith point obtained from the mocap data for the m^{th} pose, $i = 1, \ldots, l$. Then the projection of this point onto the plane corresponding to the nth view angle is given by $\mathbf{p}_{mn}^{(i)} = \mathbf{T}_n \mathbf{q}_m^{(i)}$. Here $\mathbf{T}_n$ is the transformation matrix for the n^{th} view angle which is the product of the 2×3 projection matrix and the 3×3 rotation matrix about the vertical axis. Let $\mathcal{P}_{mn} = \left(\mathbf{p}_{mn}^{(1)}, \ldots, \mathbf{p}_{mn}^{(l)}\right)$ be the tuple of 2D points

[1] http://mocap.cs.cmu.edu

representing the human figure in phase m and view n and $\mathcal{V}_{mn} = \left(\mathbf{v}_{mn}^{(1)}, \ldots, \mathbf{v}_{mn}^{(l)}\right)$ be their corresponding 2D motion vectors. Note that $\mathcal{V}$ denotes motion vectors obtained from the training data while V represents the motion vectors obtained from the test sequences. Figure 3 shows the multiple views and poses obtained from the 3D marker data. In this work we use 8 views and 8 poses.

2.2 Motion Descriptor

It is not possible to compare the sets of sparse motion vectors directly using a technique like PCA [18] because there is no ordering of the features. Instead, we aggregate the motion information in spatially local areas. Given the training data of positions $\mathcal{P}_{mn}$ and velocities $\mathcal{V}_{mn}$, we define the motion descriptor ψ_{mn} for pose m and view n as an 18-element vector containing the magnitude and phase of the weighted average motion vector in nine different spatial areas, where the weight is determined by an oriented Gaussian centered in the area. More precisely, the jth bin of the motion descriptor is given by

$$\psi_{mn}(j) = \sum_{i=1}^{l} \mathbf{v}_{mn}^{(i)} G_j(\mathbf{p}_{mn}^{(i)}), \tag{3}$$

where G_j is a 2D oriented Gaussian given by

$$G_j(\mathbf{x}) = \frac{1}{2\pi|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_j\right)^T \Sigma_j^{-1}\left(\mathbf{x}-\mu_j\right)\right), \tag{4}$$

with μ_j and Σ_j being the mean and covariance matrix of the jth Gaussian, precomputed with reference to the body center.

Figure 4 shows the nine spatial ellipses used in computing the motion descriptor, along with their Gaussian weight maps. The discriminative ability of the motion descriptor is illustrated in the rest of the figure. The confusion matrix shows the pseudo-colored Euclidean distance between the motion descriptors of all pairs of 64 configurations, with zero values along the diagonal. It is clear from this matrix that motion alone carries sufficient information to discriminate between the various poses and views in nearly all situations. The bottom row of the figure shows the descriptor bin values for two cases: three different views of the same pose, and the same view of three different poses. Because they capture the motion of the upper body, the first several bins have similar values, while the last several bins representing the lower body show a larger degree of variation. It is this larger variation in the lower part of the body that gives the descriptor its discriminatory power.

3 Pose and Viewpoint Estimation

Hidden Markov Models (HMMs) are well suited for the estimation of human gait over time. HMMs are statistical models consisting of a finite number of states which are not directly observable (hidden) and which follow a Markov chain, i.e., the likelihood of occurrence of a state at the next instant of time conditionally depends only on the

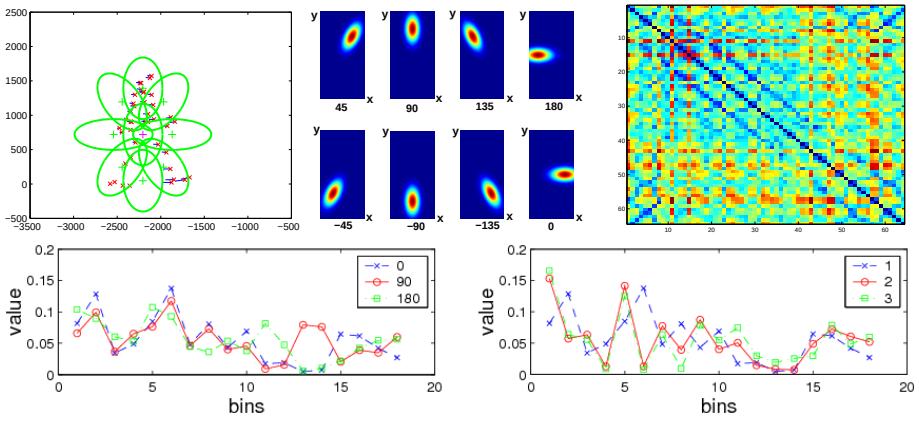

Fig. 4. TOP: The proposed motion descriptor (left), weight maps (middle) of all but the central Gaussian used for computing the motion descriptor, and the 64×64 confusion matrix (right) for 8 poses and 8 views. BOTTOM: The motion descriptor bin values for different views of the same pose (left), and for the same view of different poses (right).

current state. Each discrete pose for each viewpoint can be considered as a hidden state of the model. Assuming that the pose of a human walking is a Markov process, the observation probabilities can be computed from the image data using the motion of the limbs, and the state transition probabilities and priors can be determined beforehand. The goal is then to determine the hidden state sequence (pose estimates and viewpoint) based on a series of observations obtained from the image data.

Let $\lambda = (A, B, \pi)$ be the HMM, where A is the state transition probability matrix, B is the observational probability matrix, and π is the prior. Let the configuration c_t represent the hidden state of the model at time t, and let O_t be the observation at that time. There is a finite set of states $\mathcal{S} = \{(1,1), \ldots, (M,N)\}$ corresponding to each pose and view angle. The state transition probability is $A(i,j) = P(c_{t+1} = s_j | c_t = s_i)$, $s_i, s_j \in \mathcal{S}$, i.e., the probability of being in state s_j at time $t+1$ given that the current state is s_i. The observation probability is given by $B(j,t) = P(O_t | c_t = s_j)$, i.e., the probability of observing O_t at time t given that the current state is s_j. Given the HMM $\lambda = (A, B, \pi)$, and series of observations $\mathcal{O} = \{O_0, \ldots, O_T\}$, our goal is to find the sequence of states $\mathcal{C} = \{c_0, \ldots, c_T\}$ such that the joint probability of the observation sequence and the state sequence given the model $P(O, \mathcal{C} | \lambda)$ is maximized.

The state transition probability between two states $s_i = (m_i, n_i)$ and $s_j = (m_j, n_j)$ is predefined to be

$$p(s_j | s_i) = \begin{cases} \phi_{next} & \text{if } n_i = n_j \text{ and } m_j = m_i + 1 \\ \phi_{remain} & \text{if } n_i = n_j \text{ and } m_j = m_i \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

where $\phi_{next} = 0.51$ is the probability of transitioning to the next pose, and $\phi_{remain} = 0.43$ is the probability of remaining in the same pose. Note that, as mentioned earlier, the transition probability from one view to another view is zero, creating effectively a

disconnected HMM. The observation probability is given by a normalized exponential of the Euclidean distance between the test and training motion descriptors. The optimum state sequence $\mathcal{C}$ for the HMM is then computed using the Viterbi algorithm.

4 Experimental Results

Our approach was tested on a variety of sequences of walking humans from different viewpoints, scales, and illumination conditions. The detection of articulated bodies is performed by computing the motion descriptor to each pixel of the image at three different scales and projecting the descriptor onto a line to determine the similarity with respect to a model of human motion. A strength map is generated indicating the probability of a person being at that location and scale, and the maximum of the strength map is used as the location and scale of the target. Figure 5 shows human detection based on this procedure. The unique characteristics of human motion when compared to other motions present in natural scenes is clear from the ability of such a simple procedure to detect the people. Using only motion information, the person is correctly detected in each sequence, even when the camera is moving, because only differences between motion vectors are used. Once the person has been detected, Lucas-Kanade point features are tracked through the image sequence, and the location and scale of the person is updated using the tracked points attached to the detected target. The entire process is fully automatic.

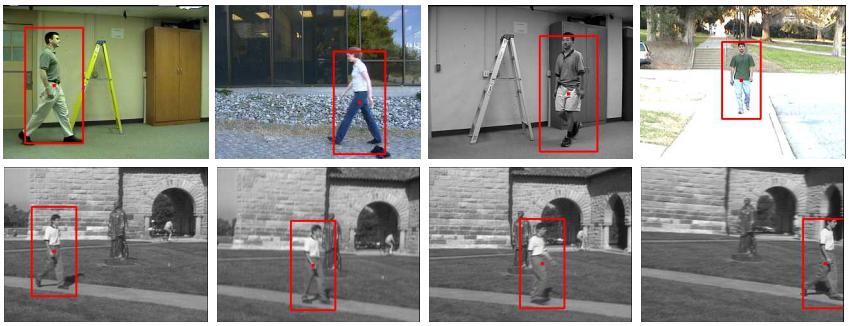

Fig. 5. Articulated motion detection for various viewpoints: right profile, left profile, at an angle, and frontal. In the bottom row, the camera is moving.

Figure 6 shows the pose estimation results for sequences captured from various viewpoints. Each sequence covers an entire gait cycle. The stick figure models correspond to the nearest configuration found in the training data by the HMM. It is important to keep in mind that point feature tracks are not very accurate in sequences such as these involving non-rigid motion and large amounts of occlusion, and a large number of point features belonging to the background cause noise in the data, especially when the camera is moving. Moreover, when the person walks toward or away from the

camera (frontal view), the pose estimation is difficult due to the ambiguity in motion. Nevertheless, the estimated poses are qualitatively correct.

The last row of the figures shows a sequence captured at night by an infrared camera. The person is wearing a special body suit fitted with reflectors that reflect the light emitted by the headlights of an oncoming vehicle. This suit has been used in psychological studies of the effectiveness of reflectors for pedestrian safety by exploiting the biomotion capabilities of the human visual system of automobile drivers [22]. The utility of a purely motion based approach can be especially seen in this sequence, in which no appearance information is available. Even without such information, the motion vectors are highly effective within the current framework for estimating the pose. To provide quantitative evaluation, Figure 7 shows the estimated knee angles at every frame of the right profile view and the frontal view sequences, along with the ground truth.

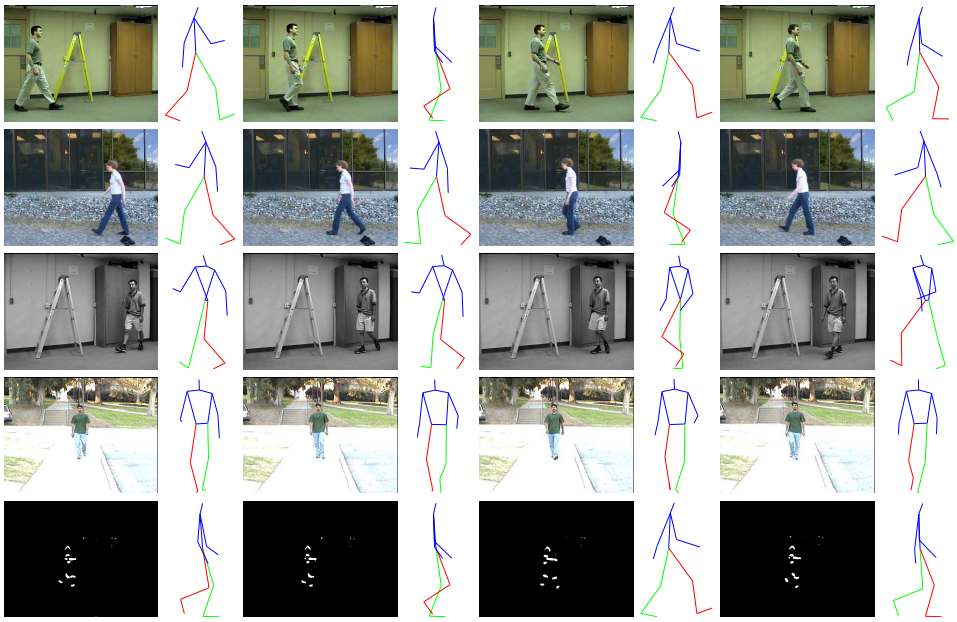

Fig. 6. Top to bottom: Pose estimation for four frames from several sequences: right profile view, left profile view, angular view, frontal view, and profile view at night with reflectors.

As can be seen from these results, our approach offers several advantages over previous motion-based approaches [3, 18, 17, 21]. First, it is invariant to scale and viewpoint, and it is able to deal with noisy video sequences captured from a moving camera. In contrast, many of the previous algorithms rely on a static camera, tightly controlled imaging conditions, and/or a particular walking direction (e.g., profile view). Another advantage of our approach is that it is easy to train, requiring only a small amount of training data since there is no need to account for all the variations in appearance that occur in real imagery. Since the estimated poses of our approach are necessarily tied

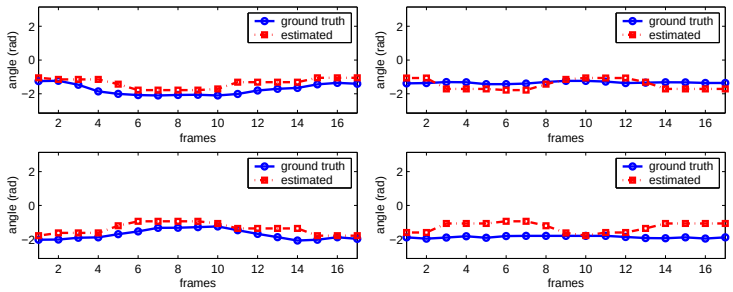

Fig. 7. Estimated and ground truth knee angles for two sequences. The top row shows the right knee, while the bottom row shows the left knee.

to the training data, it is not possible to recover arbitrary body poses not seen in the training data. Nevertheless, it may possible to train a similar detector to handle various other actions such as running or hand waving with appropriate data.

5 Conclusion

Motion is a powerful cue that can be effectively utilized for biological motion analysis. We have presented a motion-based approach for detection, tracking, and pose estimation of articulated human motion that is invariant of scale, viewpoint, illumination, and camera motion. In this spirit of one-shot learning, the approach utilizes only a small amount of training data. The spatial properties of human motion are modeled using a novel descriptor, while temporal dependency is modeled using an HMM. A clear advantage of using a purely motion based approach is demonstrated in pose estimation in nighttime sequences where no appearance information is available. In demonstrating the effectiveness of motion information alone, our intention is not to discount the importance of appearance information but rather to highlight the effectiveness of this particular cue. Future work involves exploring ways of articulated motion detection in the presence of noise, allowing the subjects to change viewpoints as they are tracked, combining the bottom-up and top-down approach for more accurate pose estimation, and incorporating appearance information for increased robustness.

Acknowledgments

We would like to thank Dr. Rick Tyrrell for graciously providing the nighttime sequence.

References

1. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception and Psychophysics **14** (1973) 201–211

2. Brostow, G.J., Cipolla, R.: Unsupervised Bayesian detection of independent motion in crowds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2006) 594–601
3. Daubney, B., Gibson, D., Campbell, N.: Real-time pose estimation of articulated objects using low-level motion. In: CVPR. (2008)
4. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: ICCV. (2003)
5. Papageorgiou, C., Poggio, T.: A trainable system for object detection. IJCV **38** (2000) 15–33
6. Wu, B., Nevatia, R.: Detection and tracking of multiple partially occluded humans by Bayesian combination of edgelet based part detectors. IJCV **75** (2007) 247–266
7. Gavrila, D., Munder, S.: Multi-cue pedestrian detection and tracking from a moving vehicle. IJCV **73** (2007) 41–59
8. Cutler, R., Davis, L.: Robust real-time periodic motion detection, analysis, and applications. PAMI **22** (2000) 781–796
9. Agarwal, A., Triggs, B.: Tracking articulated motion using a mixture of autoregressive models. In: ECCV. (2004) 54–65
10. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose limbed people. In: CVPR. (2004) 421–428
11. Urtasun, R., Fleet, D., Hertzman, A., Fua, P.: Priors for people from small training sets. In: ICCV. (2005) 403–410
12. Ramanan, D., Forsyth, D.: Finding and tracking people from bottom-up. In: CVPR. (2003) 467–474
13. Song, Y., Goncalves, L., Perona, P.: Unsupervised learning of human motion. PAMI **25** (2003) 814–827
14. Lee, M., Nevatia, R.: Human pose tracking using multiple level structured models. In: ECCV. (2006) 368–381
15. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: CVPR. (1997) 568–575
16. Lan, X., Huttenlocher, D.: A unified spatio-temporal articulated model for tracking. In: CVPR. (2004) 722–729
17. Fathi, A., Mori, G.: Human pose estimation using motion exemplars. In: ICCV. (2007)
18. Fablet, R., Black, M.: Automatic detection and tracking of human motion with a view based representation. In: ECCV. (2002) 476–491
19. Lipton, A.: Local applications of optic flow to analyse rigid versus non-rigid motion. In: ICCV Workshop on Frame-Rate Applications. (1999)
20. Niebles, J., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: CVPR. (2007)
21. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. PAMI **23** (2001) 257–267
22. Wood, J., Tyrrell, R., Carberry, T.: Unsupervised learning of human motion. Human Factors **47** (2005) 644–653