Detecting People Carrying Objects based on an Optical Flow Motion Model

Tobias Senst, Rubén Heras Evangelio and Thomas Sikora
Communication Systems Group, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin, Germany
senst,heras,sikora@nue.tu-berlin.de

Abstract

Detecting people carrying objects is a commonly formulated problem as a first step to monitor interactions between people and objects. Recent work relies on a precise foreground object segmentation, which is often difficult to achieve in video surveillance sequences due to a bad contrast of the foreground objects with the scene background, abrupt changing light conditions and small camera vibrations. In order to cope with these difficulties we propose an approach based on motion statistics. Therefore we use a Gaussian mixture motion model (GMMM) and, based on that model, we define a novel speed and direction independent motion descriptor in order to detect carried baggage as those regions not fitting in the motion description model of an average walking person. The system was tested with the public dataset PETS2006 and a more challenging dataset including abrupt lighting changes and bad color contrast and compared with existing systems, showing very promissing results.

1. Introduction

The number of video surveillance cameras is increasing notably. This leads to a growing interest in algorithms for automatically analysing the huge amount of video information generated by these devices. The development of such algorithms is being further boosted by the increasing processing power that modern CPU architectures offer. Detecting people carrying objects is a commonly formulated problem. Results can be used as a first step in order to monitor interactions between people and objects, like depositing or removing an object.

There have been several methods proposed aiming at detecting people carrying objects. The majority of these methods is based on analysing the silhouette of a person obtained by means of foreground segmentation. One of the first methods was Backpack, proposed in [10, 11]. Backpack is a two step algorithm. At first, the symmetry of the silhouette of the persons detected in a frame is analysed. Based on the assumption that the silhouette of a person is symmetric, non-symmetric parts are labelled as potential carried baggage. After that, the parts labelled in the first step showing a periodicity are discarded, assuming that these are the arms and legs.

In [1] the authors introduce a body model inspired by the human appearance and use simple constraints for the detection of carried objects. They partition the detected persons in four blocks and calculate the periodicity and the amplitude of the blocks over time. Those persons whose features do not fit the properties observed for the gait of persons walking without baggage are classified as persons carrying an object.

A blob based method is presented in [2]. The authors introduce a classification method based on k-nn classifiers using two different sets of features: foreground density features with granularity and real size features. The foreground density features are produced by dividing each tracked object into the same number of regions and calculating the proportion of foreground pixels to the total number of pixels for each of these regions.

In [12], pose preserving dynamic shape models are used to detect people carrying objects in video sequences by means of their silhouette. The authors use an iterative procedure of hole filling and outliers detection using pose preserving shape reconstruction to enhance the precision in the detection of people carrying objects.

In [21] the authors introduce a periodicity dependency pattern describing the motion of a tracked person based on the correlation of a blockwise temporal match. The intent is be independent from the results of a background subtraction approach. People carrying objects are then classified using an off-line trained support vector machine.

Recent state-of-the-art methods use the technique of temporal templates introduced by [9]. In [22] the authors use a set of Gabor based human gait appearance models on a general tensor discriminant analysis in order to detect people carrying objects. In [8] the authors use the temporal templates of people tracked in videos and match them with exemplar templates generated with a 3D Maya©

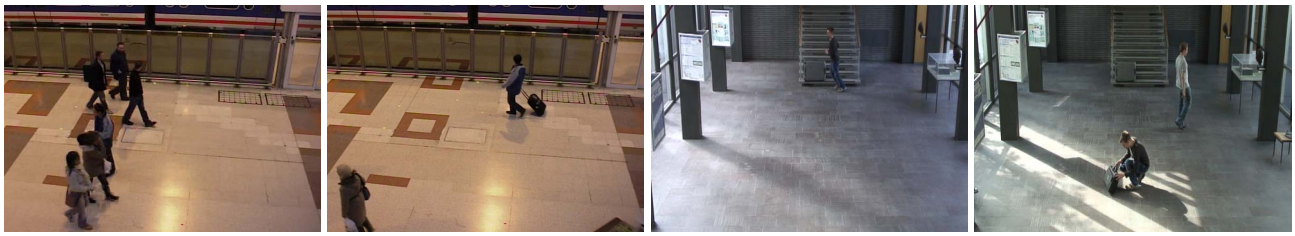


Figure 1. Two frames from each of the PETS (left) and our (right) datasets. Lighting changes are not appreciable in the PETS dataset, while in ours are drastic.

model using 8 cameras to capture the motion of different persons walking on a treadmill. In order to cope with different camera positions, scale, translation and rotation, they first decide which exemplar to use by calculating the viewing direction of the person being tracked. Candidate carried objects are then detected by the salient obtained by matching the tracked person with the best model. To enhance the detection rate, prior information and a spatial continuity assumption are used. The approach was tested on the PETS2006 public dataset.

Recent work relies on a precise foreground object segmentation obtained by means of background subtraction, which is often difficult to achieve in video surveillance sequences due to a bad contrast of the foreground objects with the scene background, abrupt changing light conditions and small camera vibrations.

In order to cope with these difficulties we propose an approach based in the motion statistics of the foreground objects. Therefore we use a Gaussian mixture motion model (GMMM) of the objects being tracked. Furthermore we propose a short-time and a long-time speed and direction independent motion descriptors in order to detect carried baggage as those regions not fitting in both motion descriptions of walking persons. The proposed method is evaluated with the PETS2006 dataset and compared to the method of Damen and Hogg. In addition, we show results obtained by using our own dataset in order to remark the improvement achieved by using optical flow information. The dataset includes more challenging surveillance scenarios with bad contrast and abrupt changing light conditions, see Figure 1.

2. Detecting People Carrying Objects

By observing the motion of people walking, we distinguish two kinds of motion: a uniform motion in the same direction as the person is moving (corresponding to the torso and the head) and a periodic motion (corresponding to the limbs). If we model the motion for an average walking person, we can detect carried objects as those motion detections not fitting in the model. Therefore, we collect statistics about the motion exhibit by pedestrians and use a novel Gaussian motion mixture model (GMMM) and motion descriptors based on the GMMM.

In order to obtain motion statistics, the method we propose starts with the bounding boxes obtained from a generic tracker [26, 18, 23]. These boxes contain the motion vectors corresponding to each pixel in the box. By comparing the statistics of the motion vectors with the main direction of the bounding box, we can differentiate pixels following uniformly the direction of the box from pixels with periodic and other kinds of motions (eg. limbs, scene background, occluded persons...). Carried baggage is then identified as uniform motion vectors pointing in the same direction as the bounding box in those pixels where no uniform motion was observed for an average walking person.

In this section we first compare the information obtained out of the optical flow computation with the information obtained from background subtraction. Then, we introduce the proposed GMMM and a motion descriptor based on this model. Finally, we show how to detect people carrying objects based on their motion description.

2.1. Motion Information obtained from Background Subtraction and Optical Flow

The most recent methods, which use motion information obtained from background subtraction or other foreground segmentation techniques, to analyse human behaviour are derived from the motion-history images (MHI) [9]. The MHI for a pixel mask $D(x, y, t)$ at time t is defined as:

$$MHI_\tau(x, y, t) = \begin{cases} \tau & , \text{if } D(x, y, t) = 1 \\ max(0, MHI_\tau(x, y, t-1) - 1) & , \text{else} \end{cases} \quad (1)$$

where τ is the maximal duration of the motion track. A benefit of MHI is the excellent run-time performance [16]. Therefore, this technique has been further exploited in many popular motion features based approaches as the modified motion-history (MMHI) [17], the motion gradient orientation (MGO) [5], the motion edge history images (MEHI) [25] or the temporal template [8]. A drawback of these methods is the dependency from background subtraction, which meets its limits in case of changing light conditions and moving platforms.

The optical flow is a representation of pixel motion in

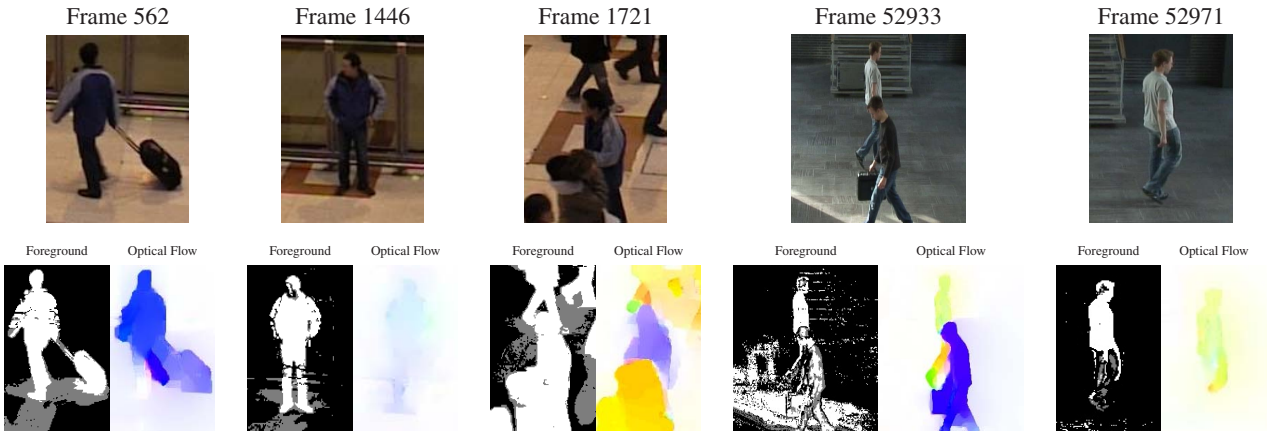


Figure 2. Examples from PETS2006 (S7-T6-B3)(frame 562 - 1721) and our dataset (frame 52933/52971) with background subtraction (white - foreground, gray - shadow, black - background) and optical flow field (color displays the angle and brighness the value -white denotes no motion-).

an image sequence. The main assumption is the *brightness constancy assumption* as stated by [19]. Thus the intensity of a small region in two consecutive images remains constant, although its position is changing. That leads to the mathematical formulation:

$$I(x, y, t) = I(x + dx\delta t, y + dy\delta t, t + \delta t), \qquad (2)$$

with $I(x, y, t)$ being the image intensity of a grayscaled image, $(dx, dy)^T$ denoting the motion vector of a point and δt a small time difference at a position $\mathbf{x} = (x, y)$. The dense set of motion vectors at time t is called optical flow. The most successful methods to solve this equation use a linearisation of Eq. 2 performed by a first order Taylor-approximation and are therefore gradient-based. This leads to an underdetermined linear system. To solve this system, two kinds of methods have been introduced containing an additional global constraint [19] or a local constraint [13] and minimizing the mean square error. Most of the state-of-the-art optical flow methods were derived from these two approaches. A local constraint implies the constancy of motion in an image region. Recent methods using local constraints are e.g. [15, 20]. Global constraints impose a condition to be fulfilled by the whole motion field. Therefore Horn/Schunck [19] introduced the smoothness constraint of the motion vectors. Recent methods use more sophisticated global constraints and robust estimators. E.g. in [6], the authors use an efficient multigrid algorithm and in [24], the Huber-L1 norm is chosen to improve the accuracy.

In the recent years, optical flow computation has becaoe more and more accurate [3]. Through the use of efficient algorithms and with the upcoming GPU computation there are a number of real-time optical flow algorithms available [15, 24].

The information obtained from the computation of the optical flow can be used in many applications [4]. In contrast to the MHI and approaches derived from it, there are few approaches using optical flow information to describe human motions. Some exceptions are those in [27, 14], which use histograms to compute feature vectors of optical flow. As an alternative to background subtraction based methods, which fail in many typical video surveillance scenarios, we use a motion model to detect pedestrians carrying objects.

In order to illustrate the difference in the information obtained out of background subtraction and optical flow computation, Figures 2 shows the results obtained for some areas containing pedestrians from the PETS2006 and our own datasets. The top row shows the tracked areas in the original frame. The bottom row shows the results of background subtraction using the method described in [28] with shadow detection as in [7], where white are foreground, gray, shadow and black, background pixels and the optical flow [24] using the colour to display the angle, and the brightness to display the value of the pixel motion (white denotes no motion).

In frame 562 we obtained good information both from optical flow and background subtraction. Frame 1446 shows the limits of optical flow based methods, which have difficultly segmenting people if they are not moving. Frame 1721 shows their advantage, which is the ability to segment people by means of their motion. In contrast, with background subtraction a big blob was obtained, but it was not possible to segment the people on it. More significant results can be observed in the results obtained for the videos of our database, where the lighting conditions change so fast that the background model fails to adapt (frame 52933) and the poor contrast of the foreground objects to the background (frame 52971) hinders good segmentation results by means of background subtraction.

To illustrate motion based models we computed the

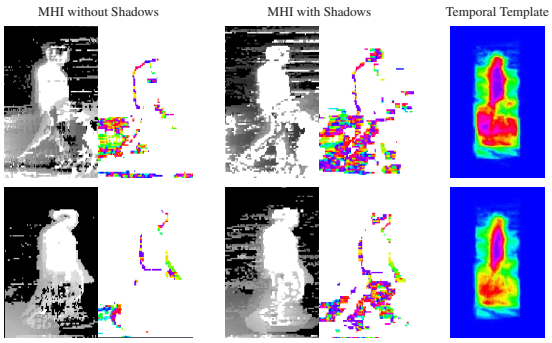


Figure 3. Motion-history images (gray) and motion-history gradient (colour coded) with and without shadow-detection and temporal templates for the frame 52933 (top) and frame 52971 (bottom) including drastic changes in illumination.

MHI, the gradient of the MHI and the temporal template used in [8] for the frames 52922 and 52971 of our dataset. Figure 3 shows the results. Gradient angles and values are represented with colour and brightness respectively. It is easy to see that the results degrade strongly with changing light conditions.

Based on this observation, we propose a statistical optical flow based motion model to describe the motion of walking people in order to classify those of them carrying objects.

2.2. Optical Flow Motion Model

The profile of a pedestrian's motion is defined through periodic and uniform motion. Periodic motion corresponds to the pendular motion of limbs and uniform motion to the torso and the head. In most cases, if people are carrying objects, a uniform motion profile not fitting to the average motion profile of a walking person can be observed.

To generate a motion model of a pedestrian we observe the motion vectors $\mathbf{v_{xy}} = (dx, dy)$ for each (x, y) position of a bounding box Ω containing a given tracked person and estimate a probability density function of the motion observed by considering the motion history in a position-wise manner. The motion history $\{\mathbf{v_{xy,1}}, \mathbf{v_{xy,2}}, \dots \mathbf{v_{xy,t}}\}$ for each position is modeled as a mixture of K Gaussians $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in a similar fashion as described in [28], with $\boldsymbol{\mu} = (\mu_{dx}, \mu_{dy})$ being the mean and $\boldsymbol{\Sigma}$ the covariance matrix. We assume that the dx and dy values are independent and have the same variances ($\boldsymbol{\Sigma} = \sigma\mathbf{I}$, with $\mathbf{I}$ being the 2x2 identity matrix) since we do not use any camera calibration information, but this could be easily incorporated into the system. This model is updated for each position $(x, y) \in \Omega$ for every new frame as described in [28]. In order to rapidly obtain a reliable motion model, we take a learning rate $\alpha_t = 1/t$ in the period $t \in [1 \dots T]$, with $T = 1/\alpha$ and α the learning rate as defined in [28]. The proposed optical flow based motion model θ_{xy} now contains a pattern of the motion observed at every position $(x, y) \in \Omega$. Figure 4 shows an example the motion model of a pedestrian of our database.

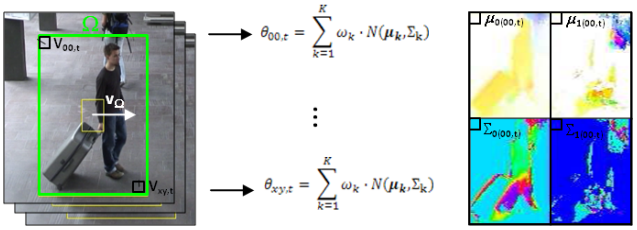


Figure 4. Pedestrian tracked with the bounding box Ω and its displacement v_Ω. The observed motion $\mathbf{v_{xy,t}}$ denoting the optical flow at the each position and are used to compute the motion statistics $theta_{xy}$ for each position. Mean μ and variance Σ are displayed colourcoded.

One of the advantages of using a GMMM to learn motion statistics is that abrupt changes of motion can be updated in an additional Gaussian, which may grow if the new direction is maintained. Thus, we obtain a much preciser and more reactive model than other approaches averaging silhouette masks over the time as [8]. The negative effect of sporadic occlusions can also be successfully prevented by using a GMMM.

2.3. Uniformity of Motion

In order to classify the motion statistic observed at each position as uniform motion or not, we calculate the probability that the motion of the bounding box $\mathbf{v_\Omega}$ is included in the motion statistic θ_{xy}:

$$p_{xy} \equiv P(\mathbf{v_\Omega}|\theta_{xy}) \tag{3}$$

The benefit of that probability p_{xy} is the independence from the speed and the direction of a tracked person. We use this probability to build a model of the short-time uniformity of motion of a pedestrian, the short-time uniformity pattern of motion. Regions corresponding to the head, torso and eventually carried baggage will have a high value, while regions corresponding to the limbs will have a lower value since they move in a pendular manner.

By collecting this information over a big sample of people carrying and not carrying objects, only the positions corresponding to the head and the torso will be reflected in the model, since baggage is usually carried at different positions. With the results of eq. 3 we estimate a probability density function ϑ_{xy} which describes the long-time uniformity of motion. Therefore a one dimensional mixture of Gaussians ($N(\mu_{p_{xy}}, \Sigma_{p_{xy}})$) is used. That is done in order to obtain a precise description of the uniformity of motion. In fact, two Gaussians are enough to obtain this density function, one to describe the main mode and the second to catch outliers. The probability density function

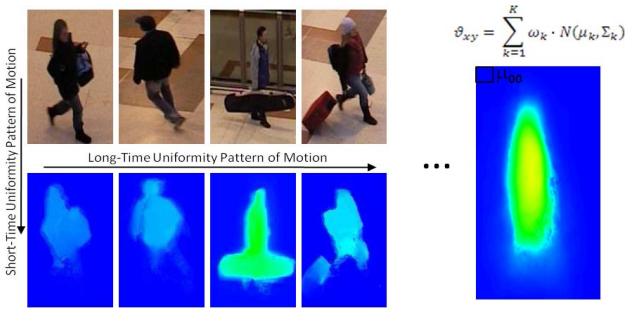


Figure 5. Examples of short-time uniformity pattern of motion (left) and long-time uniformity pattern of motion (right) obtained for the scene PETS2006.

ϑ_{xy} forms the long-term uniformity pattern of motion. Figure 5 shows the short-time uniformity of motion of some pedestrians and the long-time model obtained for the scene PETS2006 by using 106 pedestrians. Since the long-time uniformity pattern could be updated on-line, the pattern is independent from different camera viewpoints.

2.4. Classification

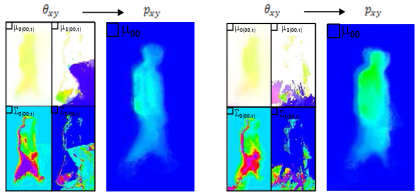
The classification process of a tracked person consists of modelling his motion with a GMMM, calculating the uniformity of motion p_{xy} that leads to the short-time uniformity pattern of motion and evaluating the membership of this value to ϑ_{xy}, obtained as in Section 2.3. Those values of p_{xy} not fitting in ϑ_{xy} are detected as potential carried baggage. The detection results are then filtered by using morphology operators, connected components analysis and a minimum area threshold.

3. Experiments and Results

To evaluate the GMMM and the derived motion description the tracked persons where annotated manually, though their corresponding bounding boxes could be easily obtained by using a general tracking method as proposed in [26]. The optical flow was computed using the GPU implementation of the Huber-L1 method proposed in [24].

At first we evaluate our method on the challenging sequences with drastic lighting changes where background subtraction based approaches fail (referring section 2.1). Figure 6 shows the obtained GMMM-based motion descriptor. The left image shows how occluding motion detections are updated in additional modes of the GMMM and thus do not perturb the motion model of the tracked person.

A numerical evaluation was done using the PETS2006 dataset and the results were compared to the results presented in [8] with a dataset containing 106 persons. The long-time uniformity pattern of motion was trained off-line but it could be trained and even updated on-line. We decided not to do this, since we wanted to strictly observe the same amount of people as [8] to obtain a fair comparison of the systems.



Frame 52933 Frame 52971

Figure 6. Results of Gaussian motion model (GMMM) on our dataset with drastic lighting changes.

Similar to [8], a detection is labelled as true if the overlap between the bounding box containing the detected object and the groundtruth box as annotated by Damen and Hogg exceeds 15% of their summed areas in more than 50% of the frames of the trajectory. The Precision-Recall is shown in Figure 7. We decided to focus our attention on the motion model. Therefore, we did not use any prior knowledge. The results obtained with our method outperformed the results presented in [8]. Figure 7 shows that using prior knowledge could increase the performance of our system additionally.

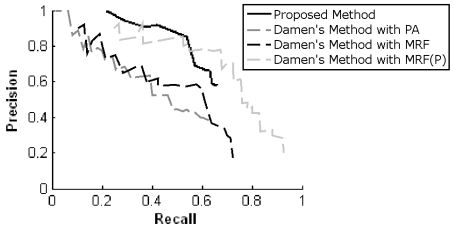


Figure 7. Comparing Precision-Recall curve for the proposed method with Damen and Hogg [8] (PA - periodicity analysis, MRF - markov random field, MRF(P) - MRF with prior knowledge).

4. Conclusions

In this paper we propose a novel method for the detection of people carrying objects based on the description of their motion. Therefore, we estimate a density function of their motion and, based on this model, we propose a motion descriptor. The fact of describing the motion independently of the speed and direction allows us to create a motion description model of an average person that can be used to detect people carrying objects. By using motion instead of colour information, the system is robust against abrupt lighting changes. The models used to describe motion can be generated on-line. The system was tested and compared with existing systems with the public dataset PETS2006 and a more challenging dataset including abrupt lighting changes and bad colour contrast, showing very promising results.

References

[1] C. B. Abdelkader and L. Davis. Detection of people carrying objects: A motion-based recognition approach. In *Automatic Face and Gesture Recognition*, pages 378–383, 2002. 1

[2] V. Atienza-Vanacloig, J. Rosell-Ortega, G. Andreu-Garcia, and J. Valiente-Gonzalez. People and luggage recognition in airport surveillance under real-time constraints. In *International Conference Pattern Recognition (ICPR 08)*, pages 1–4, 2008. 1

[3] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *ICCV*, pages 1–8, 2007. 3

[4] S. S. Beauchemin and J. L. Barron. The computation of optical flow. In *ACM Computing Surveys*, volume 27(3), pages 433466,, September 1995. 3

[5] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13:174–184, 2002. 2

[6] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnörr. Variational optical flow computation in real time. *IEEE Transactions on Image Processing*, (5):608–615, 2005. 3

[7] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003. 3

[8] D. Damen and D. Hogg. Detecting carried objects in short video sequences. In *European Conference on Computer Vision (ECCV 08)*, volume 3, pages 154–167, 2008. 1, 2, 4, 5

[9] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. In *Computer Vision and Pattern Recognition (CVPR 97)*, pages 928–934, 1997. 1, 2

[10] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis. Backpack: Detection of people carrying objects using silhouettes. *Computer Vision and Image Understanding*, 81(3):385–397, 2001. 1

[11] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000. 1

[12] C.-S. Lee and A. Elgammal. Carrying object detection using pose preserving dynamic shape models. In *Conference on Articulated Motion and Deformable Objects (AMDO 06)*, pages 315–325, 2006. 1

[13] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981. 3

[14] M. Lucena, N. de la Blanca, J. Fuertes, and M. Marn-Jimnez. Human action recognition using optical flow accumulated local histograms. In H. Araujo, A. Mendona, A. Pinho, and M. Torres, editors, *Pattern Recognition and Image Analysis*, volume 5524 of *Lecture Notes in Computer Science*, pages 32–39. Springer Berlin / Heidelberg, 2009. 3

[15] J. Marzat, Y. Dumortier, and A. Ducrot. Real-time dense and accurate parallel optical flow using cuda. In *Proceedings of the 17th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, pages 105–111, 2009. 3

[16] H. Meng, M. Freeman, N. Pears, and C. Bailey. Real-time human action recognition on an embedded, reconfigurable video processing architecture. *Journal of Real-Time Image Processing*, 3:163–176, 2008. 2

[17] T. Ogata, J. K. Tan, and S. Ishikawa. High-speed human motion recognition based on a motion history image and an eigenspace. *Transactions on Information and Systems*, 89-D(1):281–289, 2006. 2

[18] S. Pathan, A. Al-Hamadi, T. Senst, and B. Michaelis. Multi-object tracking using semantic analysis and kalman filter. In *ISPA '09: Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pages 271–276, 2009. 2

[19] B. K. P.Horn and B. G. Schunck. Determining optical flow. *Artifical Intelligence*, 17:185–203, 1981. 3

[20] T. Senst, V. Eiselein, and T. Sikora. II-LK a real-time implementation for sparse optical flow. In *International Conference on Image Analysis and Recognition (ICIAR 10)*, pages 240–249, 2010. 3

[21] T. Senst, R. Heras Evangelio, V. Eiselein, M. Pätzold, and T. Sikora. Towards detecting people carrying objects: A periodicity dependency pattern approach. In *International Conference on Computer Vision Theory and Applications (VISAPP 10)*, volume 2, pages 524–529, 2010. 1

[22] D. Tao, X. Li, X. Wu, and S. J. Maybank. Human carrying status in visual surveillance. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[23] T. Tung and T. Matsuyama. Human motion tracking using a color-based particle filter driven by optical flow. pages xx–yy, 2008. 2

[24] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-L^1 optical flow. In *British Machine Vision Conference (BMVC 09)*, 2009. 3, 5

[25] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong. Human action detection by boosting efficient motion features. In *Workshop on Video-oriented Object and Event Classification (VOEC 09)*, pages 522–529, 2009. 2

[26] J. Yu, D. Farin, and H. S. Loos. Multi-cue based visual tracking in clutter scenes with occlusions. In *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 158–163, Washington, DC, USA, 2009. IEEE Computer Society. 2, 5

[27] G. Zhu, C. Xu, W. Gao, and Q. Huang. Action recognition in broadcast tennis video using optical flow and support vector machine. In *Vision in Human-Computer Interaction workshop (HCI 06)*, pages 89–98, 2006. 3

[28] Z. Zivkovic. Improved adaptive gausian mixture model for background subtraction. In 2, editor, *International Conference on Pattern Recognition (ICPR 04)*, pages 28–31, 2004. 3, 4