TEC2011-25995 EventVideo (2012-2014)

Strategies for Object Segmentation, Detection and Tracking in Complex Environments for Event Detection in Video Surveillance and Monitoring

D3.1

PEOPLE DETECTION (IN DENSE ENVIRONMENTS)

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

AUTHOR LIST

Álvaro García Martín alvaro.garcia@uam.es

CHANGE LOG

Version	Data	Editor	Description
0.1	16-05-2014	Álvaro García Martín	Initial version
1.0	16-06-2014	José M. Martínez	First version

CONTENTS

1. Introduction

Within the computer vision field, particularly in the research area of digital image and video processing, there exists a rich variety of algorithms for foreground segmentation, object detection, event recognition, etc, which are being used in surveillance systems. People detection is one of the most challenging problems in this field. The complexity of the people detection problem is mainly based on the difficulty of modeling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of views and interactions among different people and objects. This complexity is even higher in real world scenarios such as airports, malls, etc, which often include multiple persons, multiple occlusions and background variability.

In this document, we describe the different people detection approaches that we have been working in the Video Processing and Understanding Lab in the Escuela Politécnica Superior of the Universidad Autónoma de Madrid. We firstly describe briefly the state of the art of people detection, in the second chapter we describe different people detection approaches, in the third chapter we describe two different post-processing subtask in order to improve the people detection results and finally some conclusion and future work are described in chapter 6.

1.1. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: People detection Sate of the Art
- Chapter 3: People detection approaches
- Chapter 4: People detection post-processing
- Chapter 5: Conclusions and future work

2. People detection Sate of the Art

As defined for surveillance canonical systems [1][2], every people detection approach consists mostly of, firstly, the design and training (if training is required) of a person model based on characteristic parameters (motion, dimensions, silhouette, etc) and, secondly, the adjustment of this person model to the candidates to be person in the scene. All candidates that adjust to the model will be detected or classified as person, whilst all the others will not be detected neither classified as person. Figure 1 shows the basic architecture of any people detector.

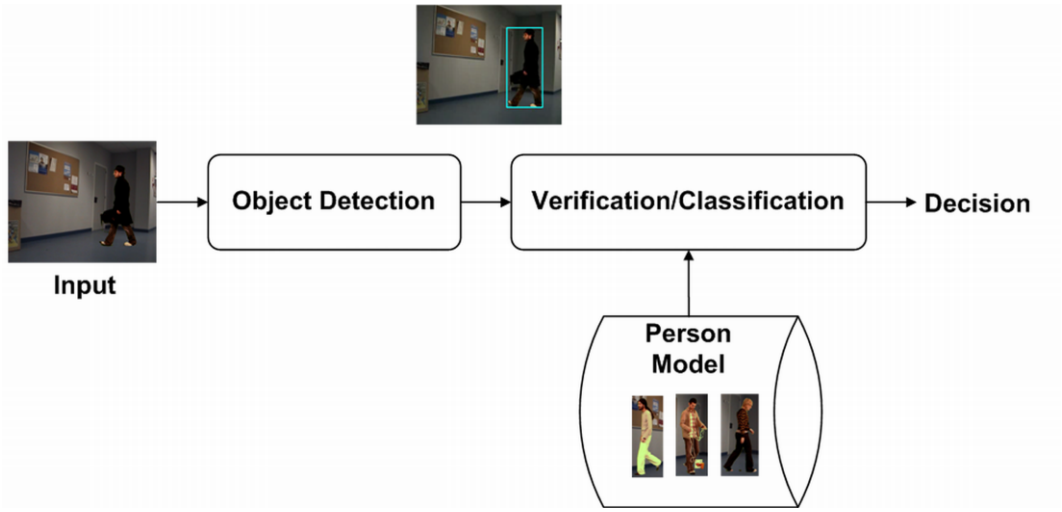


Figure 1. Canonical people detection architecture.

2.1. Object detection approach or Initial object hypotheses

There are two main conventional object detection approaches: those based on some kind of segmentation of the scene in foreground (objects) and background and those based on an exhaustive scanning approach. There are also some approaches that try to combine both approaches together. In any case, the result of this stage is the location and dimension of the different objects in the scene candidates to be a person.

2.2. Person model

As we have already commented, the verification or classification process applies a previously defined or trained person model to the objects candidates to be a person from an image or sequence and takes a final decision based on their similarity. So, the definition of a

proper person model is a critical task for the verification or classification process. There are two main discriminative information sources to characterize the people model: appearance and motion. Although the vast majority of approaches are mainly based on appearance information, there are some approaches that combine appearance and motion information in order to improve the detection results. In any case, the model should be able to discriminate between people and any other object in the scene.

3. People detection approaches

This section enumerates and describes briefly the different people detection approaches used from the state of the art: Fusion [3], Edge [4], HOG [5], ISM [6], TUD [6] and DTDP [7].

3.1. Fusion

The Fusion detector [3] is a real time detection approach based on segmentation and a holistic person model. The initial objects candidates to be person are extracted using background subtraction and the holistic person model is the combination or fusion at decision level of three simple person models: ellipse fitting [8], ghost [9] and aspect ratio.

3.2. Edge

The Edge detector [4] combines segmentation and exhaustive search in order to achieve robustness and real time operation. It is a real time adaptation of the people detection approach [11]. An individual human is modeled as an assembly of natural body parts. The main idea consists of identifying characteristic edges of each body part and generating four edge models of body parts (body, head, torso and legs). The initial objects candidates to be person are extracted using background subtraction and then those selected candidates are scanned with four independent edge feature detectors previously trained.

3.3. HOG

The HOG detector [5] is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model, evaluating different detection windows with a classifier at multiple scales and locations. The chosen person model is based on appearance information using the Histogram of Oriented Gradients.

3.4. ISM

The ISM detector [6] is a generative model for object detection and has been applied to a variety of object categories including cars, motorbikes, animals and pedestrians. The ISM people detector is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching. The chosen person model is based on appearance information using the SIFT features.

3.5. TUD

The TUD people detector [6] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original ISM detector [6] using pictorial structures. The appearance of body parts is modeled using densely sampled shape context descriptors and discriminatively trained Adaboost classifiers. As a result, it presents a strong discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts.

3.6. DTDP

The DTDP detector [7] is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original HOG detector [5]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable body part is modeled as the original HOG detector [5].

In addition, in this section we also describe two additional approaches based on the combination of appearance and motion information. The first one is a people detector based on two person models: one based on appearance and the second one based on motion, and the second approach is based on the combination of detection and tracking.

3.7. Appearance and motion

It is clear that human motion provides useful information for people detection and independent from appearance information; so [14] present an integrated system which combines an appearance people model and a motion model.

The appearance model could be anyone from the state of the art, the motion model is the IMM detector. The IMM detector is based on feature-based exhaustive. The person model is based in the characteristic movements of people using the ISM Framework [6] and the MoSIFT [13] interest points detector and descriptor. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching.

3.8. Detection and tracking

In [14], a detection/tracking collaborative scheme that integrates appearance, motion and tracking information is presented. Each task follows a parallel process and provides useful information to the other process frame by frame. The collaborative system consists of successive stages of information exchange, so the improvement introduced by one process becomes a potential self-improvement in the following stages.

The detection and tracking modules can be replaced by others without great difficulty thanks to the modular design of the system that allows a collaborative or independent performance, the generic format of the information to be exchanged (blobs and detection/tracking confidence) and the easily compatible information exchange mechanism (simple and consistent process updates).

4. People detection post-processing

Every people detector from the state of the art must maintain a balance between the number of false detections and the number of missing pedestrians. This compromise limits the global detection results. There are different post-processing subtasks in the state of the art that try to solve this problem. In this section, we describe two different people detection post-processing approaches.

4.1. People detection using people-background segmentation confidence

One of the most typical pre o post-processing approaches is the use of any kind of initial foreground/background segmentation of the scene. In this case, we describe a people detection approach that enhances people detection results making use of the information about where there are not people in the scene obtained with the people-background segmentation. The proposed filtering approach has been proposed as a post-processing, but it can be used as either a preprocessing or post-processing stage. People-background segmentation

People-background segmentation

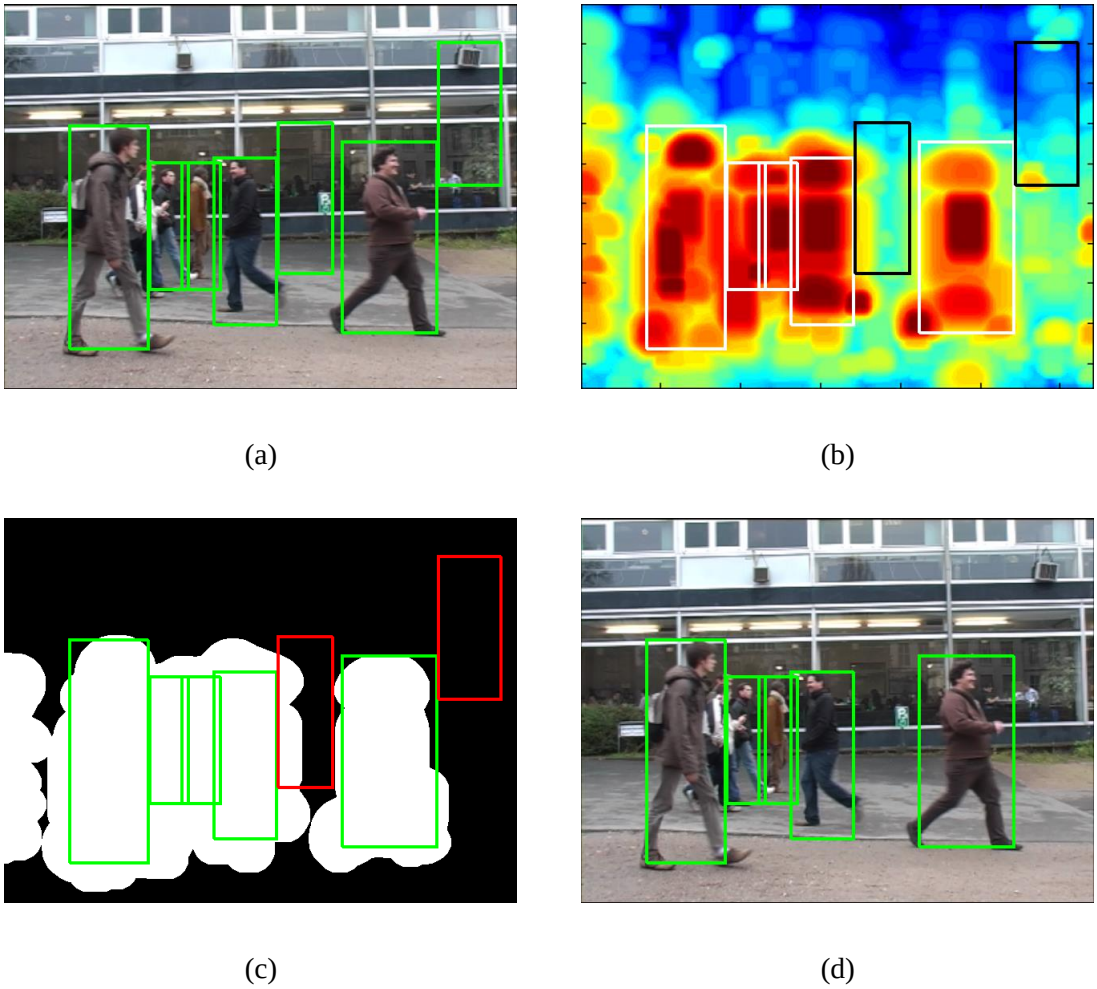
A people-background segmentation [15] is a two-class segmentation ensuring that no people or body parts are appearing in the background class. This segmentation is desirable for many computer vision applications, such as robotics and driver assistance systems. This type of segmentation is useful not only as a people detection preprocessing or post-processing step, but also for other video analysis processes such as tracking and people density estimation. While the focus of person detection approaches is to obtain a high detection performance and to reduce false positive detections, it aims at determining the areas without people in the scene by giving a higher penalty to pixels representing a person, but that have been incorrectly classified as background. This results in a segmentation mask with a bias on the background as opposed to a segmentation with bias on people.

People detection post-processing based on people-background segmentation

Firstly, people detections could be obtained using any people detector from the state of the art and the people-background segmentation is obtained as described in the previous section. Then both information sources are combined with the aim of eliminating or reducing the number of false detections while keeping, as much as possible, the number of positive detections. Since any detector consists of a list of detections in each frame. Every single detection is represented

by its position, dimensions (bounding box) and people detection confidence. The combination of human detection and people-background segmentation is made with the detections (bounding box and people detection confidence) and with the people-background confidence map (Dependent Extended Body Parts, DEBP, confidence map [15]) or the binarized and post-processed segmentation mask (Dependent Extended Body Parts Post-processed, DEBP-P, segmentation mask [15]).

Figure 2 shows one example where two false positives are eliminated using the people-background segmentation map (black blobs, Figure 2-b) and the people-background segmentation mask (red blobs, Figure 2-c).



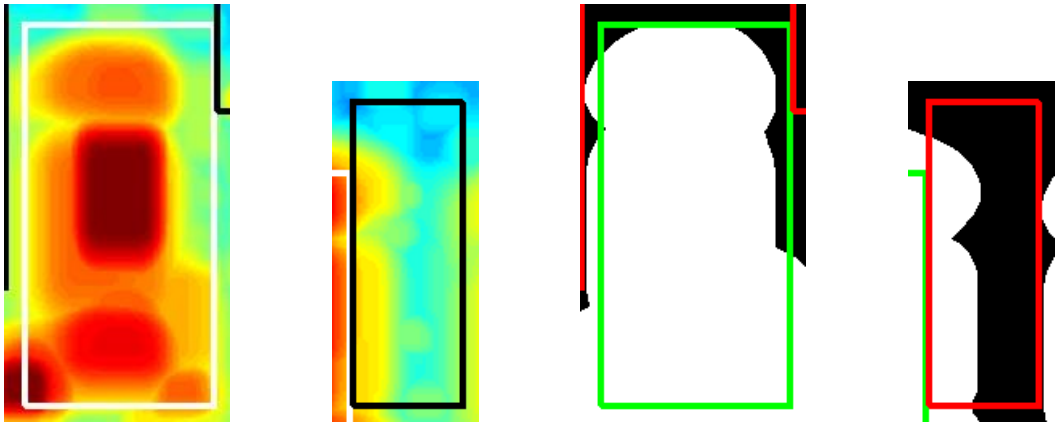
(a) (b)

(c) (d)

Figure 2. People detection system example: (a) people detections; (b) people detections over the DEBP segmentation confidence map; (c) people detections over the DEBP-P segmentation mask; and (d) final people detections.

The final people detection and segmentation confidence is the combination of the chosen people detection score or probability of being a human and the people-background segmentation

confidence. Figure 3 Shows examples over a positive and a false detection, (a) and (b) using the DEBP confidence map, (c) and (d) using the DEBP-P segmentation mask.



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

Figure 3. Examples of segmentation confidence associated with a positive and a false detection: (a) and (b) using the DEBP confidence map; (c) and (d) using the DEBP-P segmentation mask.

4.2. Decision level fusion

In this section, we describe the decision-level fusion of independent appearance based people detectors. All detectors or experts are run in parallel, and the final decision is obtained as a combination of local expert responses using fusion methods widely studied in the literature but adapted to the particular case of people detection fusion at decision-level [16]: average, product, minimum, maximum, median and majority vote.

Every people detector has their advantages and disadvantages, mainly because; each of them is based on different object extraction approaches and/or person models. The objective of this work is not to evaluate individual detectors nor to analyze the correlation among them, but to evaluate that the fusion improves results. As already commented, the main idea is the combination at decision-level of multiple people detectors from the state of the art in order to take advantage of their independent strengths and at the same time reduce their drawbacks and limitations, and therefore improve the global detection performance. As frame by frame every detector has different results, the idea consists of keeping the true positive detections that have been selected by a certain number of detectors as a true positive detection and at the same time eliminating those false positive detections that have been selected by only one detector or a small number of detectors. Figure 4 shows a visual fusion example with three detectors.

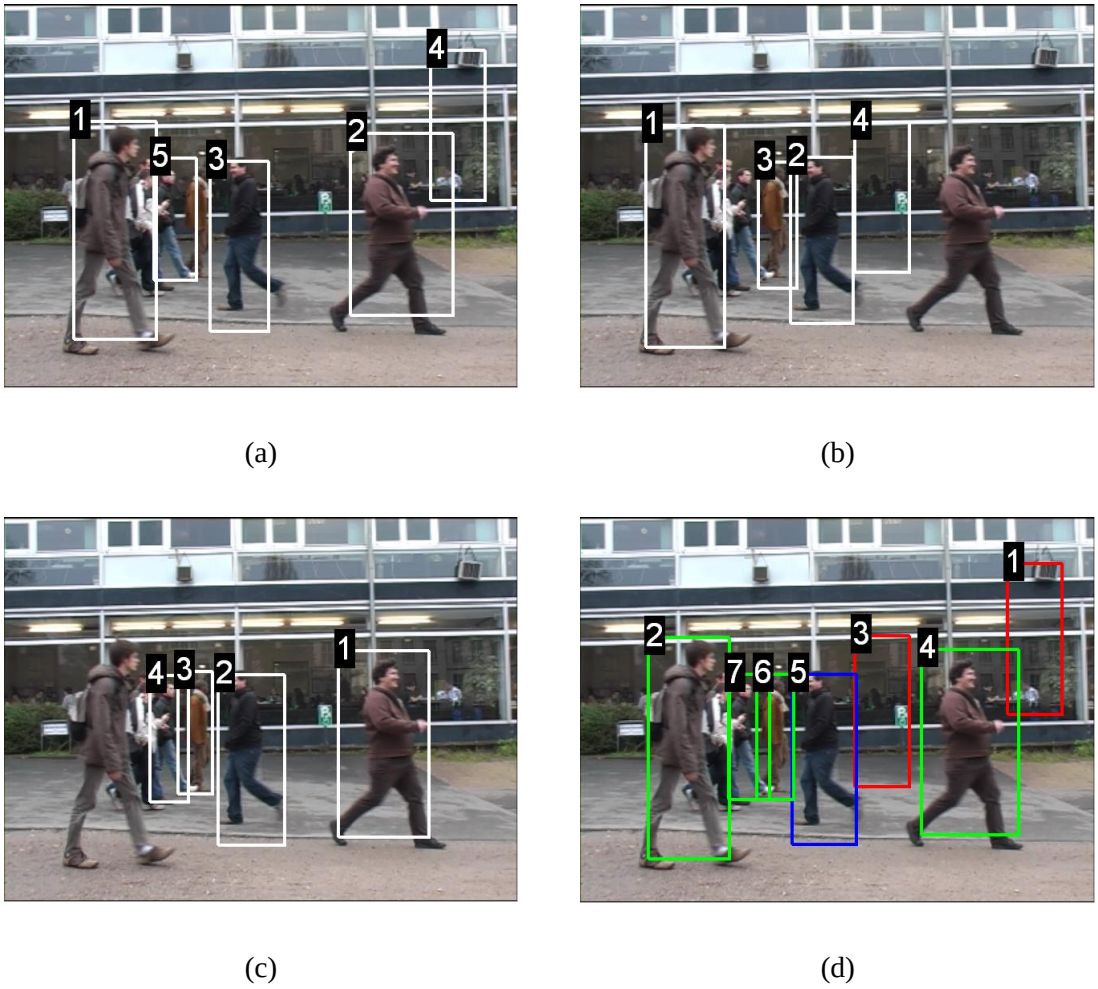


Figure 4. Visual people detection fusion example: (a) people detector outcome 1; (b) people detector outcome 2; (c) people detector outcome 3; and (d) final people detection fusion outcome.

5. Conclusions and future work

In this document, we have described the different people detection approaches that we have been working in the Video Processing and Understanding Lab in the Escuela Politécnica Superior of the Universidad Autónoma de Madrid.

We firstly have described briefly the state of the art of people detection, in the second chapter we have described the different people detection approaches that we have been working during this project, in the third chapter we have described two different post-processing subtasks in order to improve the people detection results.

People detection in real world scenarios, where the scene is totally uncontrolled, and where there are many pedestrians at the same time, is still an unsolved problem in computer vision. The use any additional information could be useful in order to improve the detection performance: using motion information (IMM, see section 3.7) or tracking approaches (Appearance+Tracking, see section 3.8).

In section 4, we address one of the main problems of people detection in video sequences; every people detector from the state of the art must maintain a balance between the number of false detections and the number of missing pedestrians. This compromise limits the global detection results. In order to reduce or relax this limitation and improve the detection results, we describe two different post-processing subtasks. Firstly, we propose the use of the people-background segmentation as a filtering stage in people detection. Then, we propose the combination of different detection approaches in order to add robustness to the detection and therefore improve the detection results.

We plan the following future research lines:

We propose the study of different fusion or combination techniques between the appearance and motion detectors, or even the creation of a single integrated Implicit Shape-Motion Model (ISMM), using the full MoSIFT description.

After showing that this People-background segmentation post-processing allows improving detection results, we propose to study the use of the people-background segmentation as a preprocessing state in order to maintain or reduce computation cost. We also propose to explore other combinations of detection and segmentation confidences.

In relation to the decision-level fusion, we propose to explore other more complex fusion possibilities, not only fixed fusion rules, but also trainable fusion rules or adaptive weights based

on online quality estimation; and not only parallel fusion schemes, but also cascade, hierarchical or hybrid.

References

[1] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," IEE Proceedings on Visual Image Signal Processing, 152(2): 192-204, 2005.

[2] Hu et al., 2004W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 34(3): 334-352, 2004.

[3] V. Fernández-Carbajales, M. A. García, and J. M. Martínez, "Robust people detection by fusion of evidence from multiple methods," In Proc. of WIAMIS 2008.

[4] A. Garcia-Martin and J. M. Martinez, "Robust real time moving people detection in surveillance scenarios," in Proc. of AVSS 2010.

[5] N. Dalal and B. Triggs, "Human detection using oriented histograms of flow and appearance," in Proc. of ECCV 2006.

[6] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in Proc. of CVPR 2005.

[7] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in Proc. of CVPR, 2009.

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

[9] F. Xu and K. Fujimura, "Human detection using depth and gray images," in Proc. of AVSS 2003.

[10] 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.

[11] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in Proc. of ICCV 2005.

[12] A. Garcia-Martin, A. Hauptmann, and J. M. Martinez, "People detection based on appearance and motion models," in Proc. of AVSS 2011.

[13] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University, 2009.

[14] A. Garcia-Martin and J. M. Martinez, "On collaborative people detection and tracking in complex scenarios," Image and Vision Computing, 30(4):345–354, 2012.

[15] A. Garcia-Martin, A. Cavallaro, and J. M. Martinez, "Peoplebackground segmentation with unequal error cost," in Proc. of ICIP 2012.

[16] L. Kuncheva. A theoretical study on six classiffier fusion strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):281-286, 2002.

Glossary

BLOB Binary Large Object

DBP Dependent Body Parts

DEBP Dependent Extended Body Parts

DEBP-P Dependent Extended Body Parts Post-processed

DTDP Discriminatively Trained Deformable Parts

HOG Histogram of Oriented Gradients

IBP Independent Body Parts

IEBP Independent Extended Body Parts

IMM Implicit Motion Model

ISM Implicit Shape Model

ISMM Implicit Shape-Motion Model

MoSIFT Motion Scale-Invariant Feature Transform

SIFT Scale-Invariant Feature Transform

TUD Technische Universität Darmstadt