An Implicit Shape model based approach to identify armed persons

Stefan Becker and Kai Jüngling

Fraunhofer IOSB
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
Gutleuthausstr. 1, 76275 Ettlingen, Germany

ABSTRACT

In addition to detecting and tracking persons via video surveillance in public spaces like airports and train stations, another important aspect of a situation analysis is the appearance of objects in the periphery of a person. Not only from a military perspective, in certain environments, an unidentified armed person can be an indicator for a potential threat. In order to become aware of an unidentified armed person and to initiate counteractive measures, the ability to identify persons carrying weapons is needed. In this paper we present a classification approach, which fits into an Implicit Shape Model (ISM) based person detection and is capable to differentiate between unarmed persons and persons in an aiming body posture. The approach relies on SIFT features and thus is completely independent of sensor-specific features which might only be perceivable in the visible spectrum. For person representation and detection, a generalized appearance codebook is used. Compared to a stand-alone person detection strategy with ISM, an additional training step is introduced that allows interpretation of a person hypothesis delivered by the ISM. During training, the codebook activations and positions of participated features are stored for the desired classes, in this case, persons in an aiming posture and unarmed persons. With the stored information, one is able to calculate weight factors for every feature participating in a person hypothesis in order to derive a specific classification model. The introduced model is validated using an infrared dataset which shows persons in aiming and non-aiming body postures from different angles.

Keywords: SIFT, ISM, object detection, weapon detection

1. INTRODUCTION

The ability to detect and track persons is one core task of many vision based surveillance and threat assessment systems and applications. In recent years, many person detection and tracking approaches have been introduced[1–3]. In most high-level applications, person detection and tracking alone is not sufficient to allow for situation assessment. Here, further information is necessary to build and maintain a meaningful picture of the environment and thus be able to assess a situation. Besides the interpretation of person trajectories which is an important part of many high-level applications, another essential part of situation assessment is the context of a person. This means, that not only the person itself is relevant in many applications, but also objects the person interacts with. Especially in military threat assessment, objects carried along by persons play an important role because they are an essential indicator for a possible threat. In particular, weapons are of interest in this context, but also other objects like bags which might contain explosives are relevant here. Thus, an important contribution to detect possible threats using vision is the detection of objects in the context of persons. This task is a very challenging one because objects which are relevant in this context are diverse. Thus, straightforward training of an object detector for a specific class is not sufficient here for two reasons: (i) object appearance might be influenced by the person which carries the object and (ii) the object itself might not be visible at all but only inferable in the context of the person. In this paper, we present an approach which copes with these challenges and uses object appearance and person context to infer the presence of an object. Specifically, our aim

Further author information:
Stefan Becker: E-mail: stefan.becker@iosb.fraunhofer.de
Kai Jüngling: E-mail: kai.juengling@iosb.fraunhofer.de

is to differentiate persons without an object from those in aiming body posture which implies the presence of a weapon. For that, we build on an Implicit Shape Model (ISM) based person detection approach introduced in[4] for infrared data. Here, a general appearance codebook is used for person representation. In this work we extend this person detection approach by introducing an additional training step which allows for detailed interpretation of a person hypothesis which was generated by the person detector. The detailed analysis allows to differentiate "normal persons" from those which might constitute a threat.

In detail, during training, which builds on samples of "normal persons" and "dangerous persons", the codebook activations and positions of participated features are stored for the relevant classes, in this case, persons in an aiming posture and unarmed persons. With the stored information, one is able to calculate weight factors for every feature participating in a person hypothesis in order to derive a specific classification model.

Only little work has been yet performed in this area. In[5] a shape analysis algorithm, combining periodic motion estimation with a static symmetry analysis of a person silhouette, has been introduced in order to determine whether a person is carrying an object under the usage of a foreground segmentation for silhouette information. Symmetric violations are interpreted as a carried object. A further development of these results is presented in[6]. Another approach is presented in[7]. Here, Branca et al. try to detect persons carrying objects such as a probe or a tin within a segmented foreground region to identify intruders in archaeological sites. The detection is based on wavelet decomposition and the classification uses a supervised three layer neural network, trained on examples of probes and tins in foreground segmentations. These approaches rely on a foreground detection that distinguishes objects and specially persons from a static background, therefore a direct comparison is not feasible.

The paper is structured as follows. In section 2 the basics of the person-detection approach is introduced and the extensions and adaptions according to[4] are detailed. In section 3, we show how the information delivered by this feature based person detector can be used to inherently classify aiming body posture in our application. The evaluation of the specific classification model is elaborated in section 4.2.

2. PERSON DETECTION

Our classification model is built on a state-of-the-art person detector presented in[4]. In this section, we briefly describe the training and detection approach and the enhancements compared to the basic trainable ISM object detection approach by[8].

2.1 Training

During the training stage, a specific object class is trained on the basis of annotated example images of the desired object category. The training is based on local features that are employed to build an appearance codebook of a specific object category.

Local features extracted from the training images on multiple scales are used to build an object category model. As local features, we picked SIFT[9] over other local features like SURF[10] or a combination of Harris[11] and Shape Context[12], because they offer robustness and other advantages as shown in[13]. The features extracted from the training images on multiple scales are used to build an object-category-model. For that, features are clustered in descriptor space to identify reoccurring features which are characteristic for the object class. To generalize from the single feature appearance and build a generic, representative object model, the clusters are represented by the cluster center. At this point, clusters with too few contributing features cannot be expected to be representative for the object category and are therefore removed from the model. The remaining feature clusters (codebook prototypes) are the basis for the generation of the ISM that describes the spatial configuration of features relative to the object center and is used to vote for object center locations in the detection process. For that, every SIFT feature found in the training data is compared to all codebook prototypes. The spatial occurrence of the feature is added to the spatial distribution of a codebook prototype if the similarity (sum of squared differences in descriptor space) of a feature and the prototype is above an assignment threshold. A weight factor which is based on the similarity of prototype and feature is used to model the importance of the feature in the spatial distribution. A single feature can contribute to more than one codebook entry (fuzzy assignment).

2.2 Detection

In order to detect objects of the trained class in an input image, SIFT features are extracted. These features are then matched with the codebook prototypes, activating prototypes with a match above a threshold t_{sim}. The spatial distribution of the prototypes casts votes for object center locations in a 3D (2 image dimension and scale) hough-voting-space. The voting space is divided into a discrete grid in x-, y-, and scale-dimension to allow fast identification of promising object hypothesis locations. Each grid that defines a voting maximum in a local neighborhood is taken to the next step, where voting maxima are refined by mean shift to accurately identify object center locations.

At this point there are two extensions in the used detector by[4] to the work of Leibe et al.[8] First, the vote weights are not equally distributed over all features and codebook entries, but assigned with a weight factor determined by the feature similarities. The weight factor $\rho(f_k, C_i)$ for an assignment of an image feature f_k and a codebook entry C_i is defined by:

$$p(C_i|f_k) = \frac{t_{Sim} - \rho(f_k, C_i)}{t_{Sim}}. \tag{1}$$

Where $\rho(f_k, C_i)$ is the sum of squared differences in descriptor space. The same distance measure is used for the weight factor $p(V_{\vec{x}}|C_i)$ of a vote for an object center location $\vec{x}$ when considering a codebook entry C_i. The vote location $\vec{x}$ is determined by the ISM learned in training. Here, $\rho(f_k, C_i)$ is the similarity between a codebook representant and a training feature that contributes to the codebook entry. The overall probability for and weight of a vote w^{Vote} is:

$$w^{Vote} = p(C_i|f_k)p(V_{\vec{x}}|C_i). \tag{2}$$

Secondly, the training data dependency is reduced. In the initial approach by Leibe et al., all votes that contributed to a maximum are used to score a hypothesis and decide which hypotheses are treated as objects and which are discarded. According to this the voting and thus the hypothesis strength depends on the amount and the character of training data. Features, that have often been seen in the training data, result in codebook entries with a large amount of contributing features and thus in a vast of votes for a single object center location with only the evidence of a single image feature. Since a feature-count independent normalization is not possible at this point, this can result in false-positive hypotheses with a high strength, generated by just a single or very few false-matching image features. Because a single image feature can only provide evidence for an object hypothesis once, just a single vote - the one with the highest similarity of image- and codebook-feature - count for image-feature/hypothesis combination. The score of a hypothesis can thus directly be inferred from the sum of all V contributing votes, without the need for a normalization:

$$\gamma_\Phi = \sum_{i=1}^{V} w_i^{Vote}. \tag{3}$$

Because objects at higher scales can be expected to generate much more features than those on lower scales this score is furthermore divided by the volume of the scale-adaptive search kernel (see[8] for details). The result of the detection step is a set of object hypotheses Φ, each annotated with a score γ_Φ. This score is subject to a further threshold application. All object hypotheses below that threshold t_{score} are removed from the detection set Φ.

3. OBJECT DETECTION IN THE PERIPHERY OF A PERSON

In this section, we present a classification approach with the key ability to differentiate between armed and unarmed persons in image data. The Implicit Shape Model (ISM) based person detection using SIFT features, outlined in section 2 is capable of handling both infrared and visible sensor data. Compared to the ISM person detection strategy, an additional training step is introduced that allows interpretation of a person hypothesis delivered by the ISM. Using trained class specific activation profiles and local distribution of features, weight

factors are calculated for every feature participating in a person hypothesis in order to derive a specific classification model. From the underlying concept, the detector described in section 2 is independent from the object category, but for the goal to detect objects in the periphery of a person, enhancements are necessary. The immediate surroundings of the object are modified through the presence of a person and often only parts of the carried object are visible. Therefore extracted features do not match with prototypes, which represent fully visible objects. Furthermore carried objects do not offer many characteristic patterns and the details of carried objects show a high similarity to background image structures. This can result in a false-hypothesis with a high score γ caused from background features. Hence, variation of the detection parameters such as t_{sim}, t_{score} can not solve the problem. According to this, the stand-alone ISM detector is only applicable in cases where the relevant object categories offer adequate characteristic structures.

A hierarchic detection, i.e. first applying a person detection with a corresponding codebook and then an object detection with a corresponding codebook, extends the number of detectable object categories only by some particular categories. Through this enhancement the stable person detector is used to eliminate false-hypotheses in the background, but only objects that are not carried too close to the person's body can be detected additionally.

To solve the problem of a modified object environment due to the presence of a person, the separation between the person and the object codebook is set off. Thus changes of local gradients are considered and as a result we receive a combined codebook that is able to deliver strong hypotheses for persons with and without carried objects. In addition to variation of shapes and view angles for an object category persons with and without objects have to be included in the ISM training data. From the detection set Φ, we do not deliver evidence if a person is carrying an object. Therefore an additional training is needed in order to derive a specific classification model.

3.1 Codebook activation

In order to get evidence if a person carries an object, it is essential to figure out which codebook prototypes include information about the presence of this object. Hence, an analysis of the activation of all prototypes of the combined codebook is accomplished to allocate the prototypes between classes. Therefore additional training images are divided into the classes "persons with a weapon in an aiming posture" and "persons without a weapon in a normal posture". The recording of the codebook activation is orientated towards the Bag-of-Words (BoW) method. BoW can be used for an image category classification, where histograms represent the quantized occurrence of a local image fragment[14]. Visual words, which are quantized vectors with attributes such as color and texture, are used for similarity calculation.

Through the codebook index of the prototypes one gains a visual vocabulary of sorts. With the BoW method one receives a representation of a special image category by a typical frequency distribution of the visual vocabulary. To inquire a class specific activation profile - assimilable to a visual word - the ISM detector is applied for all class separated training images. For a particular hypothesis the information which prototypes were activated from features contributing to a hypothesis is stored. So one receives after M ISM based person detection for the classes "person with weapon in an aiming posture" and "person without weapon in a normal posture" the mean activity $\bar{a}_i$ of a codebook entry for M positive person hypotheses for a specific class (Fig. 1).

A BoW classification of a test image is examined through the minimum distance between the specific frequency distribution for category and the frequency distribution ascertained in the test image. In comparison to BoW categorization, not all histograms or codebook entries are used to classify a test person, but exclusively activated codebook entries are incorporated. This is necessary because BoW is a statistical approach where all features extracted in an image influence the classification. The ratio of extracted features to the size of a visual word histogram differs widely to the ratio of activated codebook entries per person hypothesis to the total number of codebook entries. Since the ISM representation takes variation of shapes and view angles into account, only few prototypes can participate in a hypothesis. Therefore a consideration of features not matching with prototypes for a particular hypothesis leads to a false interpretation.

With the established activation profile, the similarity α between a person detected with a combined codebook and a class can be defined by:

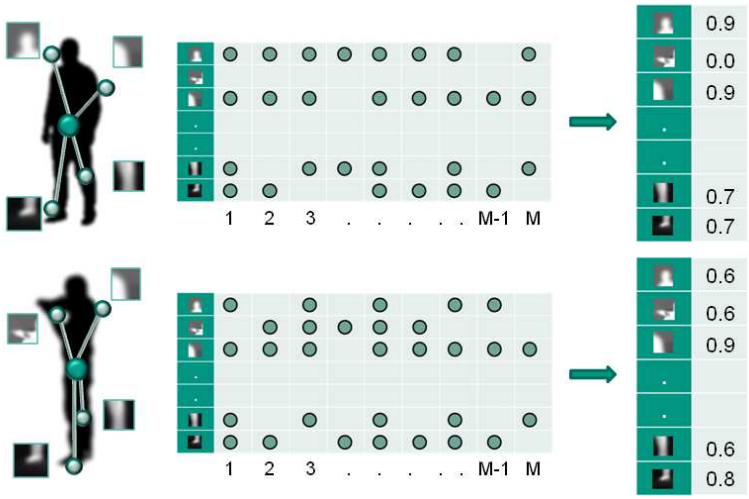

Figure 1. Visualization of the codebook activation recording. For the classes "person with weapon in an aiming posture" (down) and "person without weapon in a normal posture" (up) the mean activity $\bar{a}_i$ of a codebook entry for M positive person hypothesis is stored.

$$\alpha = \sum_{i=1}^{L} \sum_{f=1}^{F_i} (1 - (\bar{a}_{Test,f} - \bar{a}_i)). \tag{4}$$

Where $L \in \{1, .., N\}$ is the number of activated codebook entries per hypothesis, N the number of prototypes stored in the codebook and F_i the number of corresponding features per codebook entry. For a single frame interpretation $\bar{a}_{Test}$ is 1. The single terms of the activation similarity are the first weight factor of the complete classification model.

3.2 Inclusion of spatial information

One major advantage of the ISM is neglected when the codebook is interpreted as a single visual word histogram of sorts without the spatial distribution stored in the codebook. Not only the activated codebook entries are delivered by an ISM hypothesis, but additionally the position of the hypothesis center is known.

By means of the features that contribute to a hypothesis, the relative position of the keypoints to the center can be determined. But this relative position carries an uncertainty regarding the real corresponding vote position inside the Kernel. Due to the fact that a matching feature is only considered once for the calculation of the hypothesis score γ, one particular ISM offset can be attached to this feature. For scale adaption this value is divided by the scale dimension of the extracted feature. Considering scale adaption and the uncertainty of the real position inside the Kernel, the ISM offsets can be used to assign codebook entries corresponding feature positions. During the recording of the activation profiles, the ISM spatial distribution is partitioned in class specific relative codebook entry positions with scale adaption.

Based on this class specific spatial distribution, a spatial similarity can be calculated for a new person hypothesis. With the codebook indexes, a matching feature position can be compared with the estimated spatial distribution of the classes. Introducing a threshold t_{Offset} which determines the maximal distance between keypoint positions for a new hypothesis and the stored class specific positions, we can define the spatial similarity d for a class by:

$$d = \sum_{i=1}^{L} \sum_{f=1}^{F_i} \frac{t_{Offset} - min_{k=1}^{K}(\sqrt{(x_{Offset,k} - x_{Test,i,f})^2 + (y_{Offset,k} - y_{Test,i,f})^2})}{t_{Offset}}. \tag{5}$$

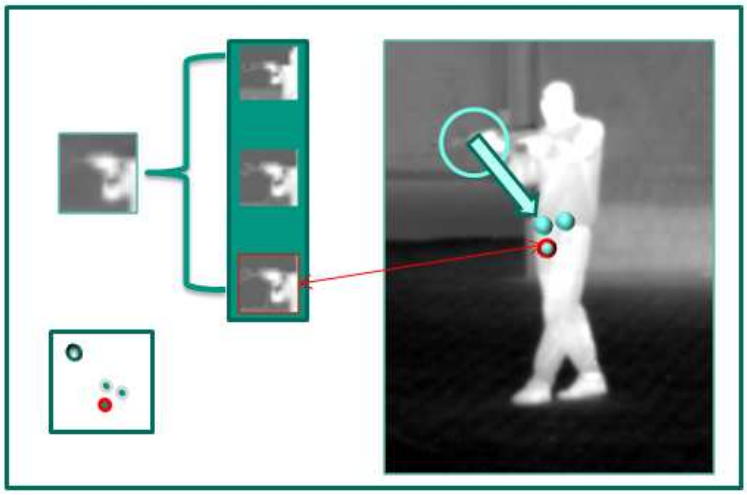

Figure 2. Features with corresponding codebook entry and spatial distribution P_C from the ISM, that were assigned through annotation to the carried object weapon (left). person detection with activated codebook entry (right). Due to the calculated vote weight one can unambiguously infer the connection to a training feature (red).

Where $K \in \{0, .., M\}$ is the number of relative positions per codebook entry stored during M training hypotheses for a class, $L \in \{1, .., N\}$ is the number of activated codebook entries per hypothesis, N the number of prototypes stored in the codebook and F_i the number of corresponding features per codebook entry. The values x_{Offset} and y_{Offset} are the class specific relative positions of a codebook entry which are determined in training and can be calculated with $(x, y)_{Offset} = \frac{\delta_{Keypoint,(x,y)} - \delta_{Vote,(x,y)}}{scale\ feature}$. Here the terms δ_{Vote} and $\delta_{Keypoint}$ are the image positions of the strongest vote and keypoint position of features resulting in a hypothesis. The values x_{Test} and y_{Test} are relative positions for the inquired hypothesis which can be calculated analogous to $(x, y)_{Offset}$. With the terms d_i one gains the second weight factor for the classification model.

3.3 Semantic annotation

Using only the activation and spatial similarities, a direct position calculation of a carried object is not feasible. Codebook entries that show a significantly stronger activity for the class with the carried object are indicative for the carried object. According to this, the keypoint position of a feature that matches with such a codebook entry can be used for the calculation of the carried object position.

Furthermore, the enhancements by[4] in the detection step of the ISM approach provide an unambiguous inference to a training feature, that has created a specific vote. Hence, only the strongest vote is considered for the calculation of the hypothesis score γ, the corresponding offset leads back to a particular training feature. During the ISM training an annotation can be added to the codebook entries for features that can be associated with carried objects. The training data carried object annotation can directly be used to annotate training-features found on carried objects with semantic identifiers. The person hypotheses resulting from detection consist of a number of votes. These were generated by specific entries (that refer to training features) in certain codebook entries that were activated by image features. Using the annotation of these entries, one is able to infer the semantics of (some) image features that contribute to a person hypothesis (see Fig. 2).

In case of a classification of an armed person, the keypoint of an activated annotated feature can be used to determine the weapon position. If more than one annotated feature is involved in a hypothesis, the vote weight factor $p(V_{\vec{x}}|C_i)$ is considered for a more detailed position calculation. For Z involved annotated features the weapon position is defined by equation:

$$(x, y)_{Objekt} = \frac{\sum_{i=1}^{Z} \delta_{Keypoint,(x_i,y_i)} \cdot p(V_{\vec{x}}|C_i)}{\sum_{i=1}^{Z} p(V_{\vec{x}}|C_i)}. \quad (6)$$

Table 1. Classification results for the classes "person with weapon in an aiming posture" and "person without weapon in a normal posture".

single frame decision		
number of persons armed/unarmed	**sensitivity**	**specificity**
100/100	0.95	0.9

Due to the fact that this semantic annotation has the weakness that the similarity between an image feature and the training feature is calculated only indirectly by the similarity between the (generalized) codebook prototypes and the image feature (see equation 1) and details of carried objects can show a strong similarity to background image structures, this annotation is not used as stand-alone carried object identifier. In case of classification of an armed person the activation of an annotated feature can be used for verification as well as for the position calculation.

Considering Fig. 2, it is clear that extracted training features of a carried weapon (left) do not include details of the weapon. Actually, the underlying keypoint positions are expected to be on the brighter body parts and not directly on the weapon itself. Therefore the term "recognition of an aiming posture" rather than "weapon detection" should be used.

4. MODEL VALIDATION

4.1 Classification Model

By combination of the similarities α and d a classification model arises which takes the spatial distribution of single prototypes as well as the ISM activation structure into consideration. For each desired class the similarity to the ISM person hypothesis is calculated. A classification if a person is carrying an object or if a person is in an aiming posture can be determined with the help of the maximum similarity c between a new hypothesis and the differentiation classes. This similarity c is defined by:

$$c = max_{j=1}^{U}(\sum_{i=1}^{L}\sum_{f=1}^{F_i}\alpha_{j,i,f}\cdot d_{j,i,f})\ ,\ L \in \{1,...,N\}. \tag{7}$$

Where U is the number of classes, N the number of prototypes stored in the codebook, F_i the number of corresponding features per codebook entry and $L \in \{1,..,N\}$ is the number of activated codebook entries per hypothesis. The term j shows which class is classified. In case of classification of an armed person the positions of annotated features are used to calculate the carried weapon position.

4.2 Results

For model validation, an infrared dataset which shows persons in aiming and non-aiming body postures from different angles is used. The image resolution is 640 × 480 pixel, where single persons appear in a size of about 50 × 180 pixel. The actual weapon size is only about 35 × 10 pixel, therefore feature descriptors extracted on the weapon always include structures from a person. This means a stand-alone ISM detection with a weapon codebook is not feasible. Stable keypoints occur preferably on brighter body parts such as the hand and not directly on the weapon itself. This is valid for all steps - for the ISM training, for the activation recording and for test images. For performance evaluation the sensitivity and specificity is considered. In this paper sensitivity is defined by the ratio of correctly classified persons in an aiming posture to the number of correctly detected persons in an aiming posture with the ISM and the specificity is defined by the ratio of correct classified persons in a non-aiming posture without weapon to the number of correct detected persons under such conditions with the ISM.

Examples of correctly classified person without weapon are shown in Fig. 4. The center of a person hypothesis is marked with the bigger green circle, all features involved in the hypothesis are marked with the smaller green circle. Table 1 contains the overall results of the model validation. The determined weapon region is accentuated by the red circle and annotated features are accentuated by a pink outline. Classification examples for armed

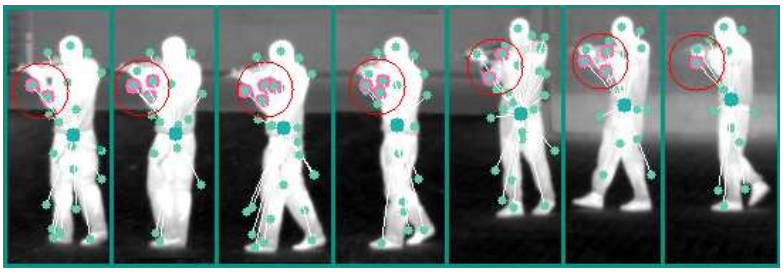

Figure 3. Classification examples for sensitivity estimation. The figure shows an armed person in an aiming posture where the weapon position is accentuated by the red circle.

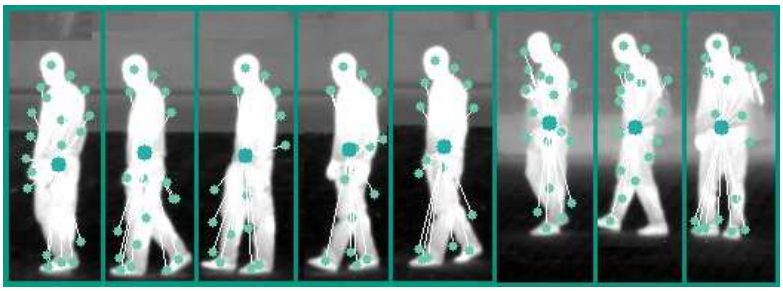

Figure 4. Classification examples for specificity estimation. The figure shows a person without weapon who is classified correctly.

persons in an aiming posture are shown in Fig. 3. In these examples, the area of the carried weapon is also calculated correctly, even though the size of the weapon is only about 35 × 10 pixel.

With the determined sensitivity of 0.95 and the specificity of 0.9 the presented classification model is capable of differentiating between unarmed persons and persons in an aiming body posture. Some of the false decisions are caused by a sub-par training. Structures from person details such as legs are not significant for the classification. Therefore prototypes of these structures must have an identical value for the mean activity for both classes. Due to a variation of shapes and view angles for persons in the training data, a small discrepancy can not be avoided. The problem can be solved by not considering all features involved in a hypothesis but only from particular regions. In this case only features above the hypothesis center are relevant for a classification. Furthermore in a single frame decision it can not be assured that in crucial regions features are extracted despite the fact that the score of the person hypothesis is strong. By extending the classification over an image sequence and using the person tracking presented in[15] the probability of this unlikely scenario can be reduced.

The presented approach can also be used for classification of persons carrying a backpack from a side view. In the free accessibly CASIA Gait Dataset C[16] a sensitivity of 0.91 and a specificity of 0.93 could be determined (see Table 2). Examples of correctly classified person with and without backpack are shown in Fig. 5 and 6. The bigger green circle is marking the center of a person hypothesis, the smaller green circles are marking features involved in a person hypothesis. Due to the extremely strong similarity between backpack details and shoulder details, the small deterioration of the performance can be explained. Here the consideration of the class specific spatial distribution has a stronger influence on the similarity c.

Table 2. Classification results for the classes "person with backpack" and "person without backpack" (CASIA Gait Dataset C).

single frame decision: CASIA Gait Dataset C		
number of persons with/without backpack	**sensitivity**	**specificity**
100/100	0.91	0.93

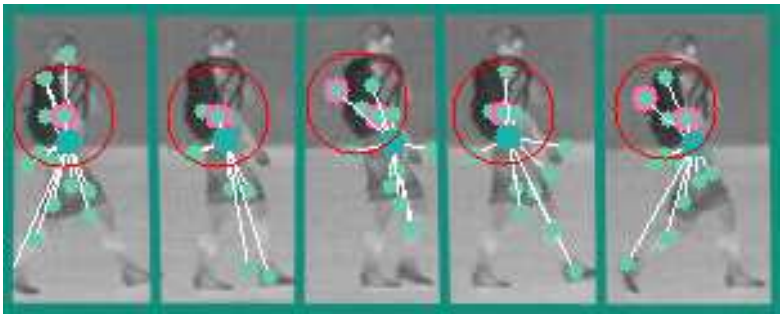

Figure 5. Classification examples for sensitivity estimation. The figure shows a person carrying a backpack where the backpack position is accentuated by the red circle.

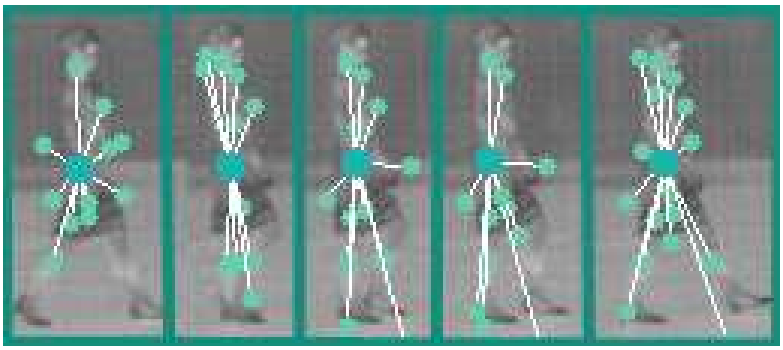

Figure 6. Classification examples for specificity estimation. The figure shows a person without backpack who is classified correctly.

5. CONCLUSION

In this paper, we present a classification approach with the key ability to differentiate between armed and unarmed persons in image data. Compared to a stand-alone person detection strategy with ISM, an additional training step is introduced that allows interpretation of a person hypothesis delivered by the ISM. Using trained class specific activation profiles and local distribution of features, weight factors or rather similarities are calculated for every feature participating in a person hypothesis in order to derive a specific classification model. The introduced classification model is evaluated in two thermal image sequences with different carried objects. The evaluation results of the test sequences show, that the classification model performs well for detection of a carried backpack in a side angle view and for identification of a person with a weapon in an aiming body posture.

REFERENCES

[1] Enzweiler, M. and Gavrila, D. M., "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(12), 2179–2195 (2009).

[2] Hu, W., Tan, T., Wang, L., and Maybank, S., "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man, and Cybernetics* **34**(3), 334–352 (2004).

[3] Yilmaz, A., Javed, O., and Shah, M., "Object tracking: A survey," *ACM Comput. Surv.* **38**(4), 13+ (2006).

[4] Jüngling, K. and Arens, M., "Feature based person detection beyond the visible spectrum," in [*Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR Workshops*], 30–37 (2009).

[5] I. Haritaoglu, R. Cutler, D. H. and Davis, L., "Backpack: detection of people carrying objects using silhouettes," in [*Proceedings of the International Conference on Computer Vision (ICCV)*], 102–107 (1999).

[6] Benabdelkader, C. and L.Davi, "Detection of people carrying objects: a motion-based recognition approach," in [*Proceedings International Coference on Automatic Face and Gesture Recognition (FGR)*], 378–384 (2002).

[7] A. Branca, M. Leo, G. A. and Distante, A., "Detection of objects carried by people," in [*Proceedings of the Conference on Image Processing (ICIP)*], 317–320 (2002).

[8] Leibe, B., Leonardis, A., and Schiele, B., "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision* **77**(1-3), 259–289 (2008).

[9] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision* **60**(2), 91–110 (2004).

[10] Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V., "Speeded-up robust features," *Computer Vision and Image Understanding* **110**(3), 346–359 (2008).

[11] Harris, C. and Stephens, M., "A combined corner and edge detection," in [*Proceedings Alvey Vision Conference*], 147–151 (1988).

[12] Belongie, S., Malik, J., and Puzicha, J., "Shape matching and object recognition using shape contexts," *Trans. on Pattern Analysis and Machine Intelligence* **24**(4), 509–522 (2002).

[13] Mikolajczyk, K. and Schmid, C., "Scale & affine invariant interest point detectors," *International Journal of Computer Vision* **60(1)**, 63–86 (2004).

[14] Bosch, A., "Image classifikation for large number of object categories," *PhD Thesis, Department of Electronics, Informatics and Automation, University of Girona* (2007).

[15] Jüngling, K. and Arens, M., "Detection and tracking of objects with direct integration of perception and expectation," in [*Proceedings Int. Conference on Computer Vision, ICCV Workshops*], 1129–1136 (2009).

[16] "Casia gait database." http://www.sinobiometrics.com.