Action recognition in the longwave infrared and the visible spectrum using Hough forests

Barbara Hilsenbeck, David Münch, Ann-Kristin Grosselfinger, Wolfgang Hübner, and Michael Arens
Fraunhofer IOSB, Ettlingen, Germany
Email: barbara.hilsenbeck@iosb.fraunhofer.de

***Abstract*—Action recognition in surveillance systems has to work 24/7 under all kinds of weather and lighting conditions. Towards this end, most action recognition systems only work in the visible spectrum which limits their general usage to daytime applications. In this work Hough forests are applied to the longwave infrared spectrum which can capture humans both in the dark and in daylight. Further, Integral Channel Features which have shown promising results in the spatial domain are applied to the spatio-temporal domain and are incorporated into the Hough forest approach. This approach is evaluated on a new outdoor dataset containing different violent and non-violent actions recorded in the visible and infrared spectrum. It is further shown that for the visible spectrum the proposed approach achieves state-of-the-art results on the KTH and i3DPost dataset.**

I. INTRODUCTION

The recognition of human actions is important for many surveillance tasks like activity monitoring or visual log-file generation ([1], [2]). Action recognition can alert operators monitoring large areas or, when no operator is present, help in the clearance of crimes by intelligent video compression and retrieval.

Most action recognition systems are based on visual data (TV), but as action recognition must work equally well under different weather and lighting conditions, it is expedient to also explore non-visible spectral bands. Infrared (IR) cameras can capture humans in low light, and further simplify the task of separating persons from background structures. Furthermore, structures on the persons clothing or other surfaces are not captured in IR which makes IR advantageous for appearance based methods. For the spatial domain Kieritz et al. showed that a person detector based on Integral Channel Features (ICFs) provides promising results in both, the visible and the infrared spectrum [3]. Action recognition systems however operate in the spatio-temporal domain. One established method for TV data is the Hough forest approach. In contrast to holistic approaches, Hough forests is based on local features and can better handle occlusions. Ciolini et al. train Hough forests with HOG-ICFs for object detection in the visible spectrum [5] and Dapogny et al. use ICFs incorporated in a pairwise conditional random forest for facial expression recognition [6]. As ICFs can be directly integrated into the Hough forest framework, it is examined whether they, apart from their successful usage in the spatial domain, are also applicable in the spatio-temporal domain. Further, this approach is applied on visible and longwave infrared (wavelength: $8 - 14\,\mu m$) data of a new action recognition dataset and differences of the classifiers are discussed.

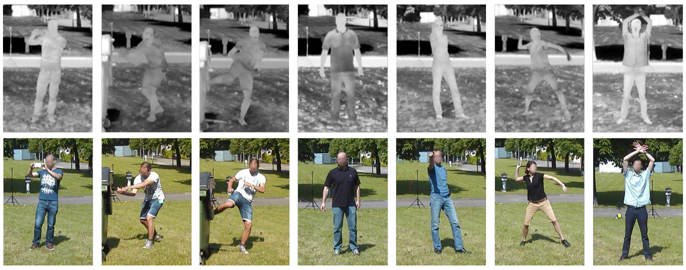

Fig. 1: The IOSB Multispectral Action Dataset. Data was recorded in the infrared (top) and visible (bottom) spectrum from a viewpoint of 0° and 90°. Ten persons performed seven actions, f.l.t.r. *film*, *hit*, *kick*, *other*, *point*, *throw*, and *wave*.

Towards this end the main contributions of this paper are: (i) Action recognition using Hough forests is performed on images recorded in the infrared and the visible spectrum. (ii) ICFs are incorporated into the approach of Hough forests. (iii) The IOSB Multispectral Action Dataset is provided, containing different violent and non-violent actions recorded from two different views with two IR and two TV cameras respectively.

II. RELATED WORK

There exist multiple established action recognition methods for the visible spectrum (see for example the surveys [7], [8]). For the infrared spectrum there is only little literature. Bhanu et al. determine 2D silhouettes in IR data and fit a kinematic 3D model to the silhouettes [9]. On this model they perform gait recognition and show experimental results for different camera views. Further, they compute Gait Energy Images (GEI) on the silhouette sequences and reduce the dimensionality by performing Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA). For classification they use a simple minimum Euclidean distance classifier and besides evaluating IR images also provide state-of-the-art results on color video data of the DARPA HumanID gait database [10]. Hossen et al. use locally adaptive regression kernels (LARK) as patch descriptors and classify different actions in the infrared spectrum using probabilistic latent semantic analysis [11]. In [12] they change their pipeline to a GEI and a naïve Bayes classifier and show results on their expanded, but not publicly available dataset. Rogitis et al. identify short-term-actions using a combination of size, speed, and appearance based features, like local binary patterns and Hidden Conditional Random

Fields [13]. By incorporating person trajectories, they recognize four clearly defined suspicious actions, e.g. *trespassing the perimeter fence and approaching the panels*. Their system is set up for an outdoor IR-based perimeter surveillance. Zhu et al. showed that the spatio-temporal features extracted from the visible and the infrared spectrum vary significantly and a direct matching is not possible [14]. Hence they studied the feasibility to adapt a support vector machine (SVM) learned from visible data to the infrared spectrum and could thereby nearly reach single-spectrum performance.

Gao et al. evaluate on an action recognition dataset, recorded in the infrared spectrum [15], which is not publicly available yet. The Infrared Action Recognition (InfAR) dataset contains different single actions and interactions captured at different times during summer and winter. The dataset is recorded at a rather small resolution of $293 \times 256\,px$. They extracted ten low-level features and trained one-vs-one SVMs using a linear and a RBF kernel. They ascertained that for IR action recognition motion information is more essential than appearance information and further that dense trajectory features can achieve the best performance among the low level features. Lee et al. recorded single frames of unsafe pedestrian behaviors at night using a far-infrared camera and proposed the pedestrians unsafe behavior (PUB) dataset [16]. Action recognition is done by combining a convolutional neural network with a boosted random forest classifier to detect unsafe behavior. As their dataset solely consists of single poses but no action cycles, it is not part of our evaluation.

Our approach differs from the mentioned approaches in that it works for both IR data and TV data. It classifies object-inherent movements without the need for background subtraction nor regarding trajectories. For evaluation we provide an outdoor action recognition dataset containing cyclic and non-cyclic action sequences recorded in IR and TV, which are potentially interesting for perimeter surveillance.

III. Augmenting Hough forests with Integral Channel Features

Our approach for action recognition is based on Hough forests [4]. Hough forests consist of a fixed set of random trees which are able to vote in the Hough space. For action recognition the Hough space encodes the hypothesis for a class located in space and time. A tree is built recursively by performing at each node a defined number of binary tests

$$t(f;p;q;\tau) = \begin{cases} 1 & \text{if } I^f(p) - I^f(q) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where I^f denotes the randomly selected feature channel, τ a randomly chosen threshold and p and q positions in the spatio-temporal feature space. The data is then split based on the test t achieving the maximum information gain. The resulting subsets are further split until a stopping criteria is met, e.g. the maximum depth of the tree or the minimum number of samples per node. The leaf nodes store the probabilities of class labels and the features displacement vectors measuring the distance to the respective action centers. As input 3D patches (e.g. of $16 \times 16 \times 5$ pixels) are used which are sampled randomly in time within the region of interest (ROI) and a randomly determined feature channel as in the original approach [4]. The ROI are designed to capture all occurring motions and have a fixed ratio of $h/w = 1.5$ with an additional border of 20% of the height of a determined bounding box where 10% are added at the top and the bottom of the bounding box respectively. The ROIs are then resized in a more manageable image format of $60 \times 40\,px$ as in [4]. For detection, densely sampled 3D patches of an unlabeled image sequence are passed through all trained trees. The probabilities from the different leaves are averaged and all votes are accumulated in the Hough space. Finally, the class label is obtained by determining the maximum peak in the Hough space.

Hough forests are extended by the use of ICFs, which have shown improved performances for pedestrian detection [3], [17]. ICFs compute the sum over local rectangular regions of different feature channels. These local sums are more robust against clutter than simple comparison on pixel level as originally proposed in [4] and since they can be efficiently computed using integral images. In the original approach ICFs included the gray channel, the three CIE-LUV color channels, gradient histograms, and the gradient magnitude channel. This approach was set up for pedestrian detection working on single 2D images. We evaluate image sequences and therefore propose the usage of a different feature set. Besides the gray image, we compute its first and second derivatives in x- and y-direction and a 9-bin HOG. As action recognition should work for different backgrounds, clothes, and skin colors we omit the color information. Further we use as motion features the TVL1 optical flow in x- and y direction. All the stated feature channels are finally filtered by a minimum and maximum filter, which results in a feature set of 48 feature channels. Especially the integration of optical flow and the maximum and minimum filter significantly improved the recognition performance in an experimental evaluation. In order to limit the search space, only ICF boxes with a fixed aspect ratio are considered. The box size is further limited to the half patch size $ICF_{width} = ICF_{height} = [0, P_w/2]$. As also box sizes of $1 \times 1\,px$ are allowed, the original approach is implicitly still comprised in the enhanced approach.

IV. The IOSB Multispectral Action Dataset

To evaluate Hough forests using ICFs on IR data, we created an outdoor database containing different violent and non-violent actions recorded at a sunny summer day from two different views from IR and TV cameras respectively [1]. For the infrared spectrum two AXIS Q1922 cameras with a spectral range of $8-14\,\mu m$ (longwave IR) and $10\,mm$ focal length for the $0°$ camera, and $35\,mm$ for the $90°$ camera, were used. For the visible spectrum, an AXIS Q5534 and an AXIS Q1755 camera were used for the $0°$ and $90°$ view respectively. The post-processed images have a resolution of $800 \times 600\,px$ for the TV and $640 \times 480\,px$ for the IR data with a frame rate of $25\,fps$.

---

[1] The dataset is publicly available under http://www.iosb.fraunhofer.de/servlet/is/49270/.

The dataset consists of ten persons (8 male, 2 female in the age of 31.2 ± 5.7) performing six different actions, namely, filming with a smartphone (*film*), hitting with a stick on an object (*hit*), kicking an object (*kick*), pointing a finger at something (*point*), throwing an object (*throw*), waving with both arms (*wave*). Furthermore, a rejection class (other) was included where the persons are standing without performing any specific action. Figure 1 shows examples of the dataset of each action class from the frontal camera view.

V. EVALUATION

The evaluation contains two parts. First, the proposed method is evaluated on the IOSB Multispectral Action Dataset, on both IR and TV data, using a ten-fold cross-validation. Second, the approach is compared with other methods on the KTH and i3DPost dataset using a five-fold and eight-fold cross-validation and the mean absolute error respectively. Further, the impact of ICFs is evaluated on the stated datasets. Although the KTH dataset is technically 'solved', we use it to illustrate the relevance of the approach on IR data and to additionally proof its usage on TV images. The parameters were selected based on experimental examination with the objective to achieve a trade off between run-time and performance: For all datasets Hough forests consisting of five trees are used. In each node of the trees 1000 tests are performed in order to select the optimal one and a minimum number of 15 patches is defined for each leaf node in order to avoid overfitting. The tree depth is set to 20, 16, and 24 for the IOSB Multispectral Action Dataset, the i3DPost and the KTH dataset respectively and the number of 3D patches sampled for each sequence are set to 4000, 2000, and 2000 for the three datasets. Table I summarizes all results performed on the IOSB Multispectral Action Dataset, which will be discussed in detail in the following sections.

A. Action recognition on IR and TV images

As it can be seen in Table I, higher recognition rates are achieved for both recorded views by evaluating images of the infrared spectrum. The performance boost through the usage of IR data in the native approach is about 6% for the 0° view and even 10% for the 90° view. Another interesting observation is that the 90° view seems more informative for action classification than the 0° view. For this action database this is reasonable since especially the *point* and *throw* actions can be classified more robust due to the outstretched arm which results in a bigger contrast to the background than observed from the front view. This effect is even stronger for the IR data as there is less background clutter and a bigger contrast between fore- and background. For other actions however a different view might be favorable.

Method	IR 0°	IR 90°	TV 0°	TV 90°
HF (ours)	87.14	95.71	81.43	84.29
HF + ICF (ours)	**95.71**	**97.14**	**88.57**	**95.71**

TABLE I: Recognition results for different IR and TV sequences of the IOSB Multispectral Action Dataset in %.

B. Incorporation of Integral Channel Features

As it can be seen in Table I, the performance increased after the incorporation of ICFs for all views. Figure 2 shows one of the two error cases, where the *point* action was confused with the *throw* action. The left image shows the Hough image for the misclassified *point* sequence. As it can be seen there are two distinct peaks, both for the *point* and the *throw* action. On the top of the right side the misclassified sequence can be seen and on the bottom the *throw* sequence of the same person. As those images even look alike for a human observer, we expect our method to struggle with this example, too.

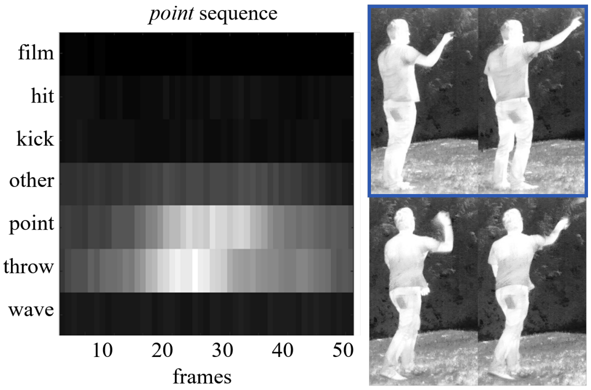

Fig. 2: (left) Hough image for an error case. (right) Sample images of the true *point* action (top), which is confused with a *throw* action (bottom).

The approach of Hough forests in combination with ICFs was further evaluated on the i3DPost and KTH dataset. For the i3DPost dataset the ROI was determined using a person detector based on boosted decision trees [3]. The recognition rates for different subsets of the i3DPost dataset can be seen in Table II. For all subsets the integration of ICFs yields a performance boost of about 4%. Whereas our approach classifies actions of only single camera views, both [18] and [19] evaluate the dataset with multi-view approaches using all camera views for training and testing. Thus, their and our methods are not that easy to compare, even though we evaluated the same dataset. Most confusion lies between the action *jump* and the actions *jump-in-place* and *run*. The actions *jump* and *jump-in-place* differ only w.r.t. a displacement in the scene and can therefore be hardly separated as our approach processes the object-inherent articulation without regarding body movement relative to the scene. The recognition rates for the KTH dataset are shown in Table III. The table is divided into three blocks based on the evaluation scheme. The first block shows results gained by a five-fold cross-validation (5F-CV), the second block shows results which were achieved using the original stated experimental setup using the data of 16 persons for training and the data of 9 persons for testing (16+9) as proposed in [20]. The authors of the third block performed a leave-one-out cross-validation (LOOCV). We used the same evaluation setup as [4] and used a five-fold cross-validation, which takes the data of 20 persons for training and 5 for testing. According

Single View approaches	10 actions	6 actions	5 actions
HF (ours)	72.81	90.36	89.06
HF + ICF (ours)	**76.72**	**94.01**	**93.12**

Multi View approaches	10 actions	6 actions	5 actions
3D-MC [18]	**80.00**	**89.58**	**97.50**
HMC [18]	76.25	85.42	95.00
Gkalelis [19]	-	-	90.00

TABLE II: Recognition results for different action sets of the i3DPost dataset in %, compared to multi-view approaches of Holte [18] and Gkalelis [19].

Method	Recognition rate
HF (ours)	93.5 (5F-CV)
HF + ICF (ours)	95.0 (5F-CV)
HF by Gall et al. [4]	93.5 (5F-CV)
Yang et al. [21]	**95.5** (5F-CV)
Liu et al. [22]	95.0 (16+9)
Sun et al. [23]	**100.0** (16+9)
Gilbert et al. [24]	95.7 (LOOCV)
Cheng et al. [25]	**97.1** (LOOCV)

TABLE III: Recognition results for the KTH dataset in %.

to [4] we limited each sequence to only one or two action cycles for training and testing and used their provided bounding boxes as ROI. The first two rows show that also for the KTH dataset a performance boost of 1.5% is reached when incorporating ICFs into the approach of Hough forests. Most confusion lies between the actions *jog* and *run* with a mean misclassification rate of 14% for the original Hough forests and 10% for the proposed approach using ICFs. The proposed method exceeds the recognition rates of [4] and yields state-of-the-art results.

VI. CONCLUSION

For surveillance tasks human action recognition is of high relevance. The demand of working 24/7 under all kinds of lighting conditions suggests the usage of IR cameras for data acquisition. Besides working at day- and nighttime, IR data has the advantage that humans can be easily detected from the background and colors of the background or the persons' clothes have no impact on the recognition result. This makes IR data especially applicative for appearance based methods. In this work action recognition using Hough forests has been performed on IR and TV data. Integral Channel Features were further incorporated in the approach leading to enhanced recognition rates for both the infrared and the visible spectrum. The approach was evaluated on the IOSB Multispectral Action Dataset containing seven violent and non-violent actions recorded in the visible and infrared spectrum and the KTH and i3DPost yielding state-of-the-art results.

REFERENCES

[1] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.

[2] D. Münch, B. Hilsenbeck, H. Kieritz, S. Becker, A.-K. Grosselfinger, W. Hübner, and M. Arens, "Detection of infrastructure manipulation with knowledge-based video surveillance," in *SPIE Security and Defence*. Int. Society for Optics and Photonics, 2016.

[3] H. Kieritz, W. Hübner, and M. Arens, "Learning transmodal person detectors from single spectral training sets," in *SPIE Security and Defence*. Int. Society for Optics and Photonics, 2013.

[4] A. Ciolini, L. Seidenari, S. Karaman, and A. D. Bimbo, "EFFICIENT HOUGH FOREST OBJECT DETECTION FOR LOW-POWER DEVICES," *IEEE Int. Conf. on Multimedia Expo Workshops*, pp. 1–6, 2015.

[5] A. Dapogny, K. Bailly, and S. Dubuisson, "Pairwise Conditional Random Forests for Facial Expression Recognition," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[6] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.

[7] D. Weinland, R. Ronfard, and E. Boyer, "A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[8] B. Bhanu and J. Han, "Kinematic-based Human Motion Analysis in Infrared Sequences," *IEEE Workshop on Applications of Computer Vision*, pp. 208–212, 2002.

[9] B. Bhanu. and J. Han, "Activity and Individual Human Recognition in Infrared Imagery," *Behavioral Biometrics for Human Identification: Intelligent Applications: Intelligent Applications*, p. 224, 2009.

[10] J. Hossen, E. Jacobs, and F. K. Chowdhury, "Human suspicious activity recognition in thermal infrared video," *SPIE Optical Engineering and Applications*, pp. 92 200E–92 200E, 2014.

[11] J. Hossen, E. L. Jacobs, and F. K. Chowdhury, "Activity Recognition in Thermal Infrared Video," *IEEE SoutheastCon*, 2015.

[12] S. Rogotis, D. Ioannidis, D. Tzovaras, and S. Likothanassis, "Suspicious activity recognition in infrared imagery using Hidden Conditional Random Fields for outdoor perimeter surveillance," *Int. Conf. on Quality Control by Artificial Vision*, pp. 95 340Q–95 340Q, 2015.

[13] Y. Zhu and G. Guo, "A Study on Visible to Infrared Action Recognition," *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 897–900, 2013.

[14] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "InfAR dataset: Infrared action recognition at different times," *Neurocomputing*, 2016.

[15] E. J. Lee, B. C. Ko, and J.-Y. Nam, "Recognizing pedestrians unsafe behaviors in far-infrared imagery at night," *Infrared Physics & Technology*, vol. 76, pp. 261–270, 2016.

[16] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough Forests for Object Detection, Tracking, and Action Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 11, pp. 2188–2202, 2011.

[17] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in *British Machine Vision Conference (BMVC)*, 2009.

[18] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3D Human Action Recognition for Multi-View Camera Systems," *IEEE Conf. on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pp. 342–349, 2011.

[19] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," *IEEE Conf. on Multimedia and Expo*, pp. 394–397, 2009.

[20] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Int. Conf. on Pattern Recognition (ICPR)*, vol. 3, pp. 32–36 Vol.3, 2004.

[21] S. Yang, C. Yuan, B. Wu, W. Hu, and F. Wang, "Multi-Feature Max-Margin Hierarchical Bayesian Model for Action Recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] L. Liu, L. Shao, X. Li, and K. Lu, "Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach," *IEEE Trans. on Cybernetics*, vol. 46, no. 1, pp. 158–170, 2016.

[23] C. Sun, I. Junejo, and H. Foroosh, "Action Recognition using Rank-1 Approximation of Joint Self-Similarity Volume," *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 1007–1012, 2011.

[24] A. Gilbert, J. Illingworth, and R. Bowden, "Action Recognition using Mined Hierarchical Compound Features," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 5, pp. 883–897, 2011.

[25] S. Cheng, J. Yang, Z. Ma, and M. Xie, "Action Recognition Based on Spatio-temporal Log-Euclidean Covariance Matrix," *Int. Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 2, pp. 95–106, 2016.