Action Recognition with HOG-OF Features*

Florian Baumann

Institut für Informationsverarbeitung,
Leibniz Universität Hannover,
{last name}@tnt.uni-hannover.de

Abstract. In this paper a simple and efficient framework for single human action recognition is proposed. In two parallel processing streams, motion information and static object appearances are gathered by introducing a frame-by-frame learning approach. For each processing stream a Random Forest classifier is separately learned. The final decision is determined by combining both probability functions. The proposed recognition system is evaluated on the KTH data set for single human action recognition with original training/testing splits and a 5-fold cross validation. The results demonstrate state-of-the-art accuracies with an overall training time of 30 seconds on a standard workstation.

1 Introduction

Human action recognition is divided into human actions, human-human interactions, human-object interactions and group activities [1]. In this paper, we address the problem of recognizing actions performed by a single person like boxing, clapping, waving and walking, running, jogging. See Figure 1.
Aggarwal and Ryoo categorize the developed methods for human activity recognition into single-layered and hierarchical approaches. These approaches are further divided into several subcategories [1]. Poppe suggests to divide the methods in global and local approaches [12]. Global approaches construe the image as a whole leading to a sensitive representation to noise, occlusions or changes in viewpoint and background. Local approaches extract regional features, leading to a more accurate representation, invariant against viewpoint and background changes.
Contribution In this paper, different feature types, such as HOG and optical flow features are used to separately learn two Random Forest classifiers which are then combined for final action recognition. HOGs are predestinated for gathering static information, since they are well-established at the task of human detection [4] while the optical flow is used for extracting motion information. Both features are local approaches, leading to an accurate representation, invariant against illumination, contrast and background changes.
Related Work Earlier work was done by Mauthner et al. [10]. The authors use HOG-descriptors for appearance and motion detection. The feature vectors are represented by NMF coefficients and concatenated. An SVM is used to learn the classifier. Similar to this work, Seo and Milanfar [16] also use an one-shot learning method. Recent

*This work has been partially funded by the ERC within the starting grant Dynamic MinVIP.

works that combine HOG- and HOF-features for single human action recognition were introduced by Wang et al. [19] and Laptev et al. [7]. Both use an SVM to recognize patterns. In contrast, we use HOG descriptors and OF trajectories to learn two independent Random Forest classifiers.

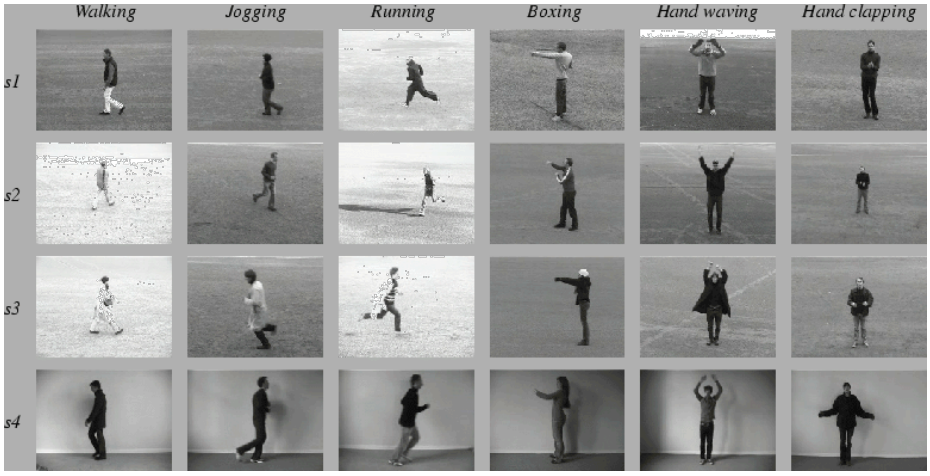


Fig. 1. Example images of KTH dataset [15]. The dataset contains six actions performed by 25 people under different conditions.

2 Approach

For human action recognition the use of static information as well as motion information is necessary to obtain robust classification results. In order to gather all information static-features from each frame and motion-features between frames are extracted. Two Random Forest classifiers are separately learned to find patterns.

2.1 Features

Frame-by-Frame Learning Static appearance information is extracted using histograms of oriented gradients. HOGs are well-established for human detection and mostly independent regarding illumination and contrast changes. First described in 2005 by Dalal und Triggs [4], HOGs became increasingly popular for different tasks of visual object detection. The computation proceeds as follows: to obtain a gradient image, a filter with [-1, 0, 1], without smoothing is applied to the image. For the computation of the HOG-Descriptor the gradient image is divided into 16x16 pixel non-overlapping blocks of four 8x8 pixel cells [4]. Next, the orientations of each cell are used for a weighted vote into 9 bins within a range of $0^\circ - 180^\circ$. Overlapping spatial

blocks are contrast normalized and concatenated to build the final descriptor.

Motion recognition The optical flow is used to gather motion information. It bases on the assumption that points at the same location have constant intensities over a short duration [6]:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t), \tag{1}$$

where $I(x, y, t)$ is the image patch, displaced in time δt with distance $(\delta x, \delta y)$ in the x-/y-direction. The proposed framework comprises the method of Lucas Kanade optical flow [8] which bases on Equation 1 but additionally assumes that the intensities are constant in a small neighborhood. To compute the optical flow descriptor, strong corners in two consecutive frames at t and $t+1$ are detected with the Shi-Tomasi corner detector [17]. Next, tracking of these feature points is realized by a pyramidal Lucas Kanade tracker [2].

2.2 Random Forest by Leo Breiman [3]

A Random Forest consists of CART-like decision trees that are independently constructed on a bootstrap sample. Compared to other ensemble learning algorithms, i.e. boosting [5] that build a flat tree structure of decision stumps, a Random Forest uses an ensemble of decision trees and is multi-class capable. A completed classifier consists of several trees $1 \leq t \leq T$ in which the class probabilities, estimated by majority voting, are used to calculate the sample's label $y(\boldsymbol{x})$ with respect to a feature vector $\boldsymbol{x}$:

$$y(\boldsymbol{x}) = \underset{c}{\operatorname{argmax}} \left(\frac{1}{T} \sum_{t=1}^{T} I_{h_t(\boldsymbol{x})=c} \right) \tag{2}$$

The decision function $h_t(\boldsymbol{x})$ provides the classification of one tree to a class c with the indicator function I:

$$I_{h_t(\boldsymbol{x})=c} = \begin{cases} 1, & h_t(\boldsymbol{x}) = c, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Classification A sample is classified by passing it down each tree until a leaf node is reached. A classification result is assigned to each leaf node and the final decision is determined by taking the class having the most votes, see Equation (2).

2.3 Combination

For each feature-type a Random Forest is separately learned to yield independent classifiers. The HOGs are used to learn a frame-by-frame classifier, so that a feature consists of a histogram obtained from each frame. A majority vote of all frames is leading to the sample's label while the class probabilities are averaged. Optical flow trajectories are calculated between two consecutive frames while a feature is constructed by concatenating the trajectories of all frames. Figure 2 shows an overview about the implemented framework. The final decision is determined by combining the class probabilities $\Pr(A_i)$ obtained by the HOG classifier and $\Pr(B_i)$ obtained by the optical flow classifier with the product law[1]: $\Pr(A_i \cap B_i) = \Pr(A_i) \Pr(B_i)$.

[1] With the assumption that events A_i and B_i are independent.

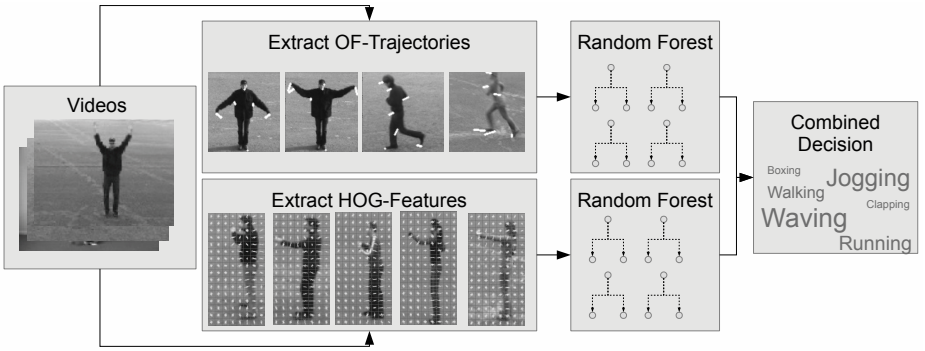


Fig. 2. Overview about the recognition system. In two parallel units static-features and motion-features are computed. For each feature-type a Random Forest classifier is separately learned. The final decision is determined by combining both classifiers using the product law.

3 Experimental Results

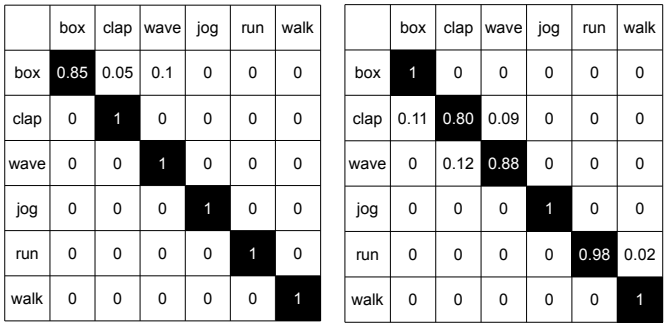
The proposed method is applied to the task of human action recognition. The goal is to recognize actions in several environments, under different illumination conditions and performed by different subjects. For the evaluation a well-established, publicly available dataset is used and the results are compared with several state-of-the-art methods.

3.1 KTH Action Dataset [15]

The KTH is a well-established, publicly available dataset for single human action recognition, consisting of 600 video files from 25 subjects performing six actions (walking, jogging, running, boxing, waving, clapping). Similar to [11], a fixed position bounding box with a temporal window of 32 frames is selected, based on annotations by Lui [9]. Presumably, a smaller number of frames is sufficient [14]. Furthermore, the original training/testing splits from [15] as well as a 5-fold cross validation strategy are used.
Jogging, running, walking, waving and clapping are perfectly learned but boxing and clapping/waving are confused. Table 1 compares the proposed framework with several state-of-the-art methods. Figure 3(a) shows the confusion matrix for 5-fold cross validation. The method achieves state-of-the-art results. Figure 3(b) shows the results for original training/testing splits. The proposed framework achieves competing results. Presumably due to the smaller training set the results are more worse than the 5-fold cross validation results. The overall training time is about 30 seconds, on a standard notebook with a single-threaded C++ implementation.

Method	Validation	Accuracy (%)
Schindler and Gool [14]	5-fold	87,98
Zhang et al. [20]	5-fold	94,60
Proposed method	**5-fold**	**96,44**
Laptev et al. [15]	Original split	91,80
Zhang et al. [20]	Original split	94,00
Wang et al. [18]	Original split	94,20
Proposed method	**Original split**	**94,31**
O'Hara and Draper [11]	Original split	97,90
Sadanand and Corso [13]	Original split	98,20

Table 1. Accuracies (%) in comparison of the proposed framework to state-of-the-art methods on the KTH dataset.



(a) (b)

Fig. 3. Confusion matrix for the KTH dataset. (a): 5-fold cross validation, (b) Original splits.

4 Conclusion

In this paper a simple and efficient framework for single human action recognition is proposed. Optical flow features are used to gather motion information between frames while static object information is extracted by using histogram of oriented gradients. With a frame-by-frame learning approach two Random Forest classifiers are separately built and the final decision is determined by combining both class probabilities. The proposed framework is evaluated using two validation strategies on the well-known, publicly available KTH dataset for single human action recognition. The results demonstrate state-of-the-art accuracies while obtaining an overall training time of 30 seconds on a standard workstation.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Computing Surveys 43(3), 16:1–16:43 (Apr 2011)
2. Bouguet, J.Y.: Pyramidal implementation of the lucas kanade feature tracker. In: Intel Corporation (2000)
3. Breiman, L.: Random forests. In: Machine Learning. vol. 45, pp. 5–32 (2001)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. vol. 1, pp. 886–893 vol. 1 (2005)
5. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Machine Learning, Proceedings of the Thirteenth International Conference on. pp. 148–156. IEEE (1996)
6. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial Intelligence 17 (1981)
7. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on (2008)
8. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence (1981)
9. Lui, Y.M., Beveridge, J., Kirby, M.: Action classification on product manifolds. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. pp. 833–839 (2010)
10. Mauthner, T., Roth, P.M., Bischof, H.: Instant action recognition. In: Proceedings of the 16th Scandinavian Conference on Image Analysis. pp. 1–10. SCIA '09, Springer-Verlag, Berlin, Heidelberg (2009)
11. O'Hara, S., Draper, B.: Scalable action recognition with a subspace forest. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. pp. 1210–1217 (2012)
12. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976 – 990 (2010)
13. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on (2012)
14. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. pp. 1–8 (2008)
15. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Pattern Recognition. (ICPR). Proceedings of the 17th International Conference on. vol. 3, pp. 32–36 Vol.3 (2004)
16. Seo, H.J., Milanfar, P.: Action recognition from one example. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(5), 867–882 (May)
17. Shi, J., Tomasi, C.: Good features to track. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. pp. 593–600. IEEE (1994)
18. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition, (CVPR). IEEE Conference on. pp. 3169–3176 (2011)
19. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference, (BMVC). (2009)
20. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision , (ECCV). vol. 7574, pp. 707–721 (2012)