Adv. Radio Sci., 11, 101–105, 2013
www.adv-radio-sci.net/11/101/2013/
doi:10.5194/ars-11-101-2013

Advances in Radio Science

Improving pedestrian detection using MPEG-7 descriptors

H. Lietz[1], **M. Ritter**[2], **R. Manthey**[2], **and G. Wanielik**[1]

[1]Chemnitz University of Technology, Professorship on Communications Engineering, Reichenhainer Straße 70, 09126 Chemnitz, Germany
[2]Chemnitz University of Technology, Chair Media Informatics, Straße der Nationen 62, 09111 Chemnitz, Germany

Correspondence to: H. Lietz (holger.lietz@etit.tu-chemnitz.de)

Abstract. During the last decade, modern Pedestrian Detection Systems made massive use of the steadily growing numbers of high-performance image acquisition sensors. Within our naturalistic driving environment, a lot of different and heterogeneous scenes occur that are caused by varying illumination and weather conditions. Unfortunately, current systems do not work properly under these hardened conditions. The aim of this article is to investigate and evaluate observed video scenes from an open source dataset by using various image features in order to create a basis for robust and more accurate object detection.

1 Introduction

Due to the increasing interest of research in the field of Vulnerable Road User (VRU) protection, a series of Pedestrian Detection Systems (PDS) using in-vehicle sensors has been developed within the past years. A state-of-the-art overview on different PDS as well as an evaluation of those systems is given in (Dollár et al., 2011). Most of the current systems use Histograms of Oriented Gradient (HOG) features, since they yield good results in the field of pedestrian detection, but fail completely in some situations. Therefore, many approaches use additional features like Haar Wavelets (Papageorgiou et al., 1998) or Local Binary Patterns (Harwood et al., 1995) in order to improve the detection and to compensate the weaknesses of single feature systems. In this contribution, we present a novel multi-feature approach which combines HOG and MPEG-7 features.

This paper is structured into five sections. The following Sect. 2 presents a brief background of HOGs and MPEG-7 features. Section 3 introduces the details of the proposed multi-feature approach. Section 4 presents the details and results of our experiments on pedestrian datasets. The paper ends in Sect. 5 with conclusions and recommendations to further work.

2 Features

In this section, we present our multi-feature approach that uses HOGs and features of the MPEG-7 Homogeneous Texture Descriptor.

2.1 Histograms of Oriented Gradients

The Histogram of Oriented Gradients was firstly described by Lowe (2004) and employed for pedestrian detection by Dalal and Triggs (2005). Many authors of the approaches mentioned in Dollár et al. (2011) used HOGs. According to Dalal's approach, the HOG feature vector is computed by dividing the image into several sub-regions: the HOG cells and HOG blocks, whereas each HOG block consists of four HOG cells. For each cell, a normalized n-bin-histogram of weighted gradient directions is computed. The main conception of the HOGs is that objects within an image can be described by the distribution of intensity gradients and edge directions. These features correlate much stronger to the contour of an object than to its color. The big advantage of the HOG features is its translation and illumination invariance.

2.2 The MPEG-7 Homogeneous Texture Descriptor

The MPEG-7 standard or ISO 15938 (Manjunath et al., 2002) is a multimedia content description standard and was developed by the Moving Pictures Experts Group (MPEG). It does not deal with encoding of video or audio data like MPEG-1, MPEG-2 or MPEG-4, but the standard offers a framework that includes functions for annotation and retrieval of multimedia data. MPEG-7 includes standardized sets of description schemes and descriptors for audio, visual and

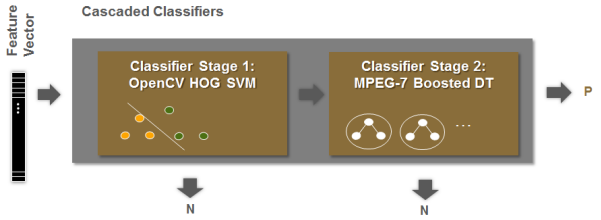

Fig. 1. Our cascaded classifiers try to reach high detection rates with the HOG-SVM in the first stage, and a second stage which focused on the avoidance of false detections, using MPEG-7 features and a boosted decision tree.

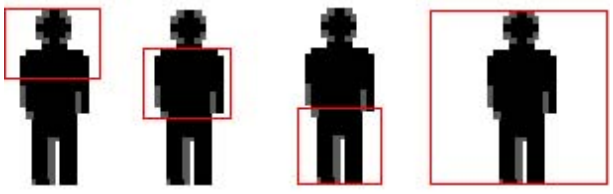

Fig. 2. MPEG-7 features are extracted from four quadratic regions (head & shoulders, torso, legs and whole ROI).

multimedia description, that have been originally composed for application in image retrieval including color, texture, shape, and motion descriptors. MPEG-7 feature descriptors have already been successfully applied in terms of human body posture estimation (Moghaddam and Piccardi, 2010) as well as visual surveillance (Annesley and Orwell, 2006). Within our evaluation, we use the features of the Homogeneous Texture Descriptor (HTD) (Ro et al., 2001) which is one part of the MPEG-7 texture descriptors for classification. It provides a homogenous image characterization for similarity retrieval based on local spatial-frequency statistics. The descriptor yields 62 integer values on every sub-window. The first two features are the mean and the standard deviation of the image (cf. Wu et al., 2001). The remaining 60 are the energy and energy deviation of the Gabor filtered responses of the 30 so-called channels in the subdivision layout of the frequency domain. For these features, the Radon transform followed by an 1-D fourier transform is applied.

3 Combined pedestrian detection system

The previously introduced features are employed in the classification algorithms which are presented in the following paragraphs.

3.1 System overview

Our approach operates in two stages: in the first stage, we use the HOG-SVM pedestrian detector of Dalal and Triggs (2005). The aim of our approach is to take benefit of the high detection rates of the HOG-SVM, but to avoid false detections. Thus, we use in the second stage a trained classifier verifying the detected regions using MPEG-7 texture features. The result of this combination is a classifier cascade that operates as follows: At first, a HOG feature vector is computed from a Region of Interest (ROI) of the input image and classified with the trained SVM. If the SVM output is non-pedestrian (N), then the output of the classifier cascade is non-pedestrian, too. If the SVM output is positive, then a new vector of MPEG-7 features of this ROI is computed and classified with a boosted decision tree. The final classifier cascade output is equal to the output of a boosted decision tree. In summary, this means that an ROI is only classified as pedestrian (P) when the output of the HOG-SVM and the boosted decision tree is pedestrian. The classification scheme is illustrated in Fig. 1.

Table 1. Classifier accuracy of different feature combinations with performance at the cross-validation (CV) and at the dedicated test set (col. @TS).

Features	# Feat.	CV Ø Acc.	CV Rank	@TS Ø Acc.	@TS Rank
HOG	2016	80.43	4	73.60	1
MPEG-7	248	89.67	1	67.89	4
MPEG-7/HOG	2264	87.21	2	70.80	2
Haar	370	84.66	3	69.55	3
LBP	1860	65.76	5	62.58	5

3.2 Classifiers and training

In the first stage of our combined classifiers, we use an implementation of the HOG-SVM, which is publicly available in the OpenCV framework, to detect persons in an input image. We used the default people detector that is trained with pedestrian and non-pedestrian images from the INRIA pedestrian dataset (Dalal and Triggs, 2005). Dalal's HOG-SVM is a powerful classifier yielding high hit rates and good classification results. However, there are still lots of false detections when applied to video sequences. The dimensions of the trained classifier are 128×64 pixel. In the second stage of our system, we trained a boosted decision tree using MPEG-7 features acquired from pedestrian and non-pedestrian images of different pedestrian datasets. We extracted features from the MPEG-7 Homogeneous Texture Descriptors from the whole ROI as well as from three equally overlapping quadratic sub-windows (head & shoulders, torso, and legs) on the vertical axis by making use of the MPEG-7 low-level feature extraction library from Başţan et al. (2010). Due to the restrictions of the employed framework, it is necessary to scale every sub-window to a size of 128×128 pixels to work properly. The regions that are used for MPEG-7 feature generation are illustrated in Fig. 2. With

Fig. 3. Five positive (left) and negative (right) samples from each dataset: INRIA (top), DaimlerChrysler (middle) and TUD-Brussels (bottom).

62 features per sub-region, we get 248 features from the Homogenous Texture Descriptor in total.

4 Experiments

To test our the MPEG-7 features, we conducted two experiments. In the first experiment, we analyzed the classification performance of those features in comparison of other well-known features. In the second experiment, we tested the benefit of the combined classifiers of the HOG-SVM and the boosted MPEG-7 decision tree. Due to different optimized versions of the feature extraction methods, runtime comparison is neglected in the following section.

4.1 Classification performances of different features

In our first experiment, we compared the classification performance of MPEG-7 features with HOGs, Haar Wavelets and Local Binary Patterns using the following publicly available pedestrian datasets: INRIA (Dalal and Triggs, 2005), DaimlerChrysler (Munder and Gavrila, 2006), and TUD Brussels (Dollár et al., 2011), illustrated in Fig. 3. Each of these datasets has annotations for pedestrian positions in images, but the DaimlerChrysler is the only dataset that has dedicated negative (non-pedestrian) examples. For the INRIA and TUD Brussels datasets, we generated negative examples by extracting ROIs at random positions in the images without any pedestrians. For each dataset, we extracted HOG, Haar, and MPEG-7 features and trained a boosted decision tree. We measured the classification accuracy firstly using a three-fold cross-validation (CV) and secondly on the dedicated test set (TS) of the applied datasets.

The mean values of the three cross-validation results and of the three dedicated test set runs (one for each dataset) are used to determine the evaluation results which are presented in Table 1.

4.2 Cascaded classifiers

In our second experiment, we applied the trained cascaded classifiers, described in Sect. 3, to a youtube video. For the evaluation we used a scenario of the driver‘s cab of a tram[1]. We manually annotated over 2,500 ROIs of pedestrians in this video with dimensions between 15×30 and 147×294 pixels. The aspect ratio was always $1:2$. The ground truth annotations only show persons being less than 50 % occluded by other persons or objects. We measured the classification performance by generating receiver operating characteristic (ROC) curves. We used the evaluation methodology of Dollár et al. (2011) which is based on the scheme laid out in the PASCAL object detection challenges (Ponce et al., 2006): a detected bounding box (BB_{dt}) and a ground truth bounding box (BB_{gt}) form a potential match, if the area of overlap is greater than 50 %.

$$a_0 = \frac{\text{area}(BB_{dt} \cap BB_{gt})}{\text{area}(BB_{dt} \cup BB_{gt})} > 0.5 \quad (1)$$

[1]Source: http://www.youtube.com/watch?v=QGLpaMpQGOA.

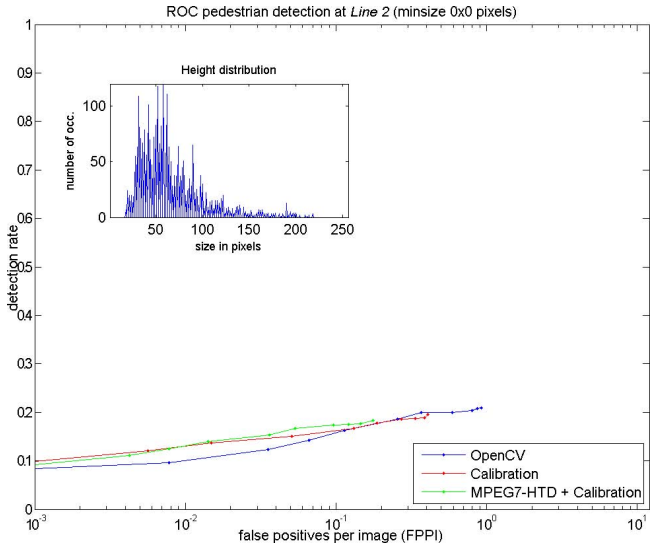

Fig. 4. ROC curve of false positives per image rates of the pedestrian detector at minimum size of pedestrian without size constraints.

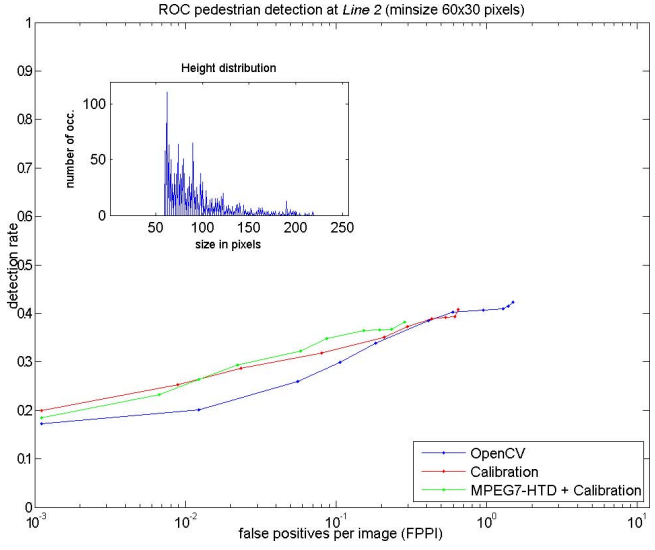

Fig. 5. ROC curve of false positives per image rates of the pedestrian detector at minimum size of pedestrian having 30 pixel in width and 60 pixel in heigth.

The results of the classification performance are illustrated in Figs. 4–6. The ROC curves were determined for three different classifiers:

1. OpenCV classiffiers applied on full-size images.

2. OpenCV classiffiers applied on image regions in that most of the pedestrians appear.

3. A cascaded classiffier applied on image regions in that most of the pedestrians appear.

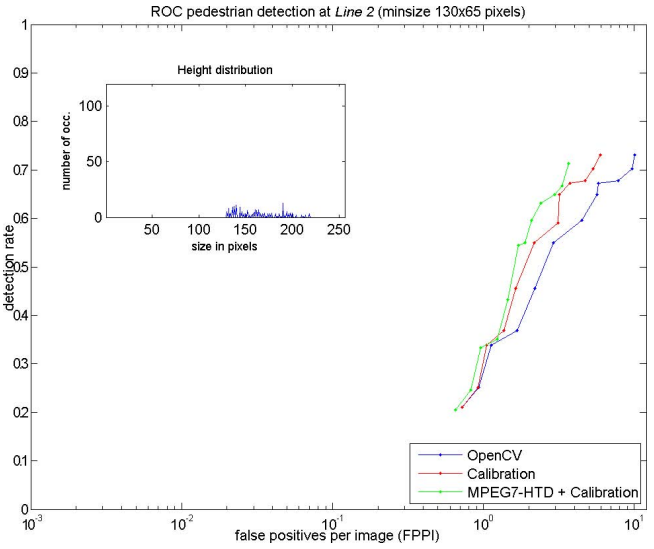

Fig. 6. ROC curve of false positives per image rates of the pedestrian detector at minimum size of pedestrian having 65 pixel in width and 130 pixel in heigth.

In cases 2 and 3, we assumed that pedestrians move on a flat ground plane. Thus, we ignored detections obviously located in the sky. By using an accurate camera calibration, human intelligence can be integrated into the detection system and the number of false detections can be reduced significantly. Because of the limitation of detection size of the OpenCV HOG-SVM, our evaluation is performed on unconstraint sizes (Fig. 4) as well as on sizes 60×30, and 130×65 pixel (Figs. 5 and 6). Due to the size constraints, the number of occurences of pedestrian ROIs in the ground trouth changes in the three different situations. The histogram of the pedestrian heights in pixels is shown in the upper left corner of each figure.

4.3 Analysis of results and conclusion

The results of our first experiment (classification performance of different features) show that MPEG-7 features ranked first place in the cross-validation test, whereas HOG features achieved the first rank in classification accuracy on the dedicated testset. The combination of HOG and MPEG-7 features reached the second place in both categories. As one can see, the results of the cross-validation tend to be significantly better than on the dedicated testsets.

On the ROC curves from our second experiment, one can see that the single HOG classifier is clearly inferior to the cascaded classifier in any operation point. The results indicate that the detection rate is much higher when ROIs are bigger in size. The ROC curves of the smaller ROIs (Figs. 4 and 5) furthermore show similar false positive rates in contrast to the bigger ROIs of Fig. 6.

The results of the conducted experiments indicate that the combination of MPEG-7 Homogeneous Texture Descriptors and HOG features can increase the classification performance by reducing the false positive rate of the existing OpenCV classifier.

5 Summary and future work

In this contribution, we presented a novel multi-feature approach by combining HOG features and features extracted from MPEG-7 texture descriptors. We conducted two experiments in order to determine the classification performance of the single feature classifier as well as the combined system.

In addition, we showed the dependency of the size of the ROI to the detection rate and the false positive rate of the classifier. We found out that it can significantly improve classification results of existing HOG-based approaches.

In future work, we will involve the color features of the MPEG-7 standard into our evaluation, as well as further classifiers and evaluation scenarios altogether with runtime comparison analysis.

Acknowledgements. This work was partially accomplished within the project ValidAX – Validation of the AMOPA and XTRIEVAL framework (Project VIP0044), funded by the Federal Ministry of Education and Research (Bundesministerium für Wissenschaft und Forschung), Germany.

References

Annesley, J. and Orwell, J.: On the Use of MPEG-7 for Visual Surveillance, in: Proceedings of 6th IEEE International Workshop on Visual Surveillance, 2006.

Baştan, M., Çam, H., Güdükbay, U., and Ulusoy, Ö.: BilVideo-7: An MPEG-7-Compatible Video Indexing and Retrieval System, IEEE MultiMedia, 17, 62–73, 2010.

Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, in: International Conference on Computer Vision and Pattern Recognition, vol. 2, 886–893, 2005.

Dollár, P., Wojek, C., Schiele, B., and Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, 743–761, 2011.

Harwood, D., Ojala, T., Pietikäinen, M., Kelman, S., and Davis, L.: Texture Classification by Center-symmetric Auto-correlation using Kullback Discrimination of Distributions, Pattern Recognition Letters, 16, 1–10, 1995.

Lowe, D. G.: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60, 91–110, 2004.

Manjunath, B., Salembier, P., and Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface, John Wiley & Sons, 2002.

Moghaddam, Z. and Piccardi, M.: Human action recognition with MPEG-7 descriptors and architectures, in: Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams, 63–68, 2010.

Munder, S. and Gavrila, D. M.: An experimental study on pedestrian classification, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, 1863–1868, 2006.

Papageorgiou, C., Oren, M., and Poggio, T.: A general framework for object detection, in: International Conference on Computer Vision, 555–562, 1998.

Ponce, J., Berg, T., Everingham, M., Forsyth, D., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B., Torralba, A., Williams, C., Zhang, J., and Zisserman, A.: Dataset Issues in Object Recognition, in: Toward Category-Level Object Recognition, vol. 4170, 29–48, 2006.

Ro, Y. M., Kim, M., Kang, H. K., Manjunath, B., and Kim, J.: MPEG-7 Homogeneous Texture Descriptor, ETRI Journal, 2001.

Wu, P., Ro, Y. M., Won, C. S., and Choi, Y.: Texture Descriptors in MPEG-7, LNCS Computer Analysis of Images and Patterns, 2124, 2001.