Color Contribution to Part-Based Person Detection in Different Types of Scenarios

Rao Muhammad Anwer, David Vázquez, and Antonio M. López

Computer Vision Center and Computer Science Dpt.,
Universitat Autònoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{muhammad,david.vazquez,antonio}@cvc.uab.es – www.cvc.uab.es/adas

Abstract. Camera-based person detection is of paramount interest due to its potential applications. The task is difficult because the great variety of backgrounds (scenarios, illumination) in which persons are present, as well as their intra-class variability (pose, clothe, occlusion). In fact, the class *person* is one of the included in the popular PASCAL visual object classes (VOC) challenge. A breakthrough for this challenge, regarding *person* detection, is due to Felzenszwalb *et al.* These authors proposed a part-based detector that relies on histograms of oriented gradients (HOG) and latent support vector machines (LatSVM) to learn a model of the whole human body and its constitutive parts, as well as their relative position. Since the approach of Felzenszwalb *et al.* appeared new variants have been proposed, usually giving rise to more complex models. In this paper, we focus on an issue that has not attracted sufficient interest up to now. In particular, we refer to the fact that HOG is usually computed from RGB color space, but other possibilities exist and deserve the corresponding investigation. In this paper we challenge RGB space with the opponent color space (OPP), which is inspired in the human vision system. We will compute the HOG on top of OPP, then we train and test the part-based human classifier by Felzenszwalb *et al.* using PASCAL VOC challenge protocols and *person* database. Our experiments demonstrate that OPP outperforms RGB. We also investigate possible differences among types of scenarios: indoor, urban and countryside. Interestingly, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

1 Introduction

Camera-based person detection is of great interest for applications in the fields of content management, video-surveillance and driver assistance. Person detection is difficult because the great variety of backgrounds (scenarios, illumination) in which persons are present, as well as their intra-class variability (pose, clothe, occlusion). Currently, discriminative part-based approaches [1, 2], that heavily rely on dynamic part detection, constitute the state of the art for detecting persons.

Fig. 1. Annotation enrichment for PASCAL VOC 2007 dataset. First, second and third rows show images that we have annotated as *indoor*, *urban* and *countryside*, resp.

The part-based human detectors generally use the histograms of oriented gradients (HOG) introduced in [3] by Dalal *et al.* as low-level features for building person models. HOG features are computed on top of RGB color space. On the other hand, in the context of image categorization [4] it has been demonstrated the usefulness of the so-called *opponent color space* (OPP) when working with the so-called SIFT descriptor [5]. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the part-based person detection method described in [2]. Moreover, we are interested in assessing if person detection performance can be affected by the type of scenario where it is performed. In other words, we want to perform a scenario-based comparison between the OPP and RGB color spaces, when *pugged-in* for HOG-part-based person detection.

As scenarios we have chosen three relevant types: indoor, countryside and urban. In order to conduct our experiments we use the class *person* included in the popular PASCAL visual object classes (VOC) challenge [6]. We have enriched the annotation with the *indoor, countryside* and *urban* labels, both for training and testing data (Fig. 1). As we will see, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

The rest of the paper is organized as follows. In section 2 we summarize our proposal of using opponent colors with part-based person detection. Section 3 details the conducted experiments, while in Sect. 4 we discuss the obtained results. Finally, Sect. 5 draws the conclusions and future work.

2 Part-based person detector based on opponent colors

The part-based paradigm, introduced by Fischler and Elshlager dates back to 1973 [7]. It provides an elegant way of representing an object category and is particularly efficient for object localization. This model has been built and extended in many direction according to different problems in the computer vision field. Here, we will briefly overview the main principles of part-based methods.

In part-based models, the focus remains on modelling an object as having a number of parts arranged in a deformable configuration. Each part captures the appearance of the object at local level and there is some flexibility in object-parts placement to account for global deformations. The best configuration of such a model is framed on an image as an energy minimization problem which measures the score for each part and deformation score for each pair of connected parts. Part-based models can be separated into many categories depending upon the connection structure to represent the parts: constellation model, star-shaped, tree-shaped, bag of features, etc. Recently, [1, 2] has adopted the star-structured part-based model, which has shown to provide excellent results on human detection [6]. The appearance of an object is represented by histograms of oriented gradients (HOG) features in a 31-dimensional feature vector. HOG of part filters are captured at twice the resolution of the root (full-body) filter to model appearance at multiple scales. Here we follow the implementation associated to [2], whose code has been kindly put publicly available by the authors.

In this implementation, and many others derived from it, HOG features are computed on top of RGB color space. Or more precisely, on top of the *max-gradient* operation on RGB color space (*i.e.*, $\max\{\nabla R, \nabla G, \nabla B\}$). This way of computing HOG derives from the original work by Dalal *et al.* [3] where HOG features were defined in the context of a holistic person detector.

However, in the context of image categorization [4] it has been demonstrated the usefulness of the so-called *opponent color space* (OPP) when working with the so-called SIFT descriptor [5]. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the part-based person detection method in [2]. Accordingly, we briefly summarize OPP in the rest of the section.

Opponent process theory postulates that yellow-blue and red-green information is represented by two parallel channels in the visual system that combine cone signals differently. It is now accepted that at an early stage in the red-green opponent pathway, signals from L and M cones are opposed and, in the yellow-blue pathway, signals from S cones oppose a combined signal from L and M cones [8]. In addition, there is a third luminance or achromatic mechanisms in which retinal ganglion cells receive L- and M- cone input. Thus, L, M and S belong to a first layer of the retina whereas luminance and opponent colors belong to a second layer of it, forming the basis of chromatic input to the primary visual cortex. Note also that this mechanism is not random since human color vision evolved for increasing the probability of subsistence [9].

	Training			Testing	
	Windows (+)	Images (-)	Initial Windows (-)	Images	Windows (+)
Indoor	(45.5%) 4268	(36.0%) 516	(36.0%) 103200	(41.0%) 2031	(49.1%) 2252
Countryside	(18.8%) 1762	(29.0%) 414	(29.0%) 82800	(29.5%) 1463	(22.2%) 1004
Urban	(35.7%) 3350	(35.0%) 501	(35.0%) 100200	(29.5%) 1458	(28.1%) 1272
Overall	9380	1431	286200	4952	4528

Table 1. Training and testing numbers per scenario: person windows (+); images without persons (-); initial background windows (-) after sampling 200 one per image without persons; number of images for testing as well as persons to be detected.

Seeing the RGB space used for codifying color in digital images as the LMS color space of the first layer of human retina, we can also compute an opponent colors (OPP) space as follows [4]:

$$\begin{aligned} \text{red-green} &: O_1 = (R-G)/\sqrt{2} \ , \\ \text{yellow-blue} &: O_2 = ((R+G)-2B)/\sqrt{6} \ , \\ \text{luminance} &: O_3 = (R+G+B)/\sqrt{3} \ , \end{aligned} \tag{1}$$

3 Experiments

In this paper we want to address the following specific questions in the context of part-based person detection: (**Q1.**) *if our detector must work in specific scenarios, is it better to use OPP or RGB?*. This is useful to know it for specific systems that must work in specific locations (e.g., intruder detection, pedestrian detection, etc.) rather than as general computer vision systems. (**Q2.**) *if we don't know a prior the scenario in which our detector must work, is it better to use OPP or RGB?*. This question is more related to general systems that must work in a broad spectrum of environments (e.g., automatically detecting people for focusing before a camera shot).

In order to answer **Q1** we will run experiments where person classifiers, based on RGB and OPP, are trained and tested in specific scenarios. We have selected three different and relevant scenarios: indoor, countryside and urban (Fig. 1). In particular we will run the part-based method summarized in Sect. 2, with the only difference of the input color space used before computing the HOG descriptors: we run equivalent experiments for RGB and OPP. We will use the *person* class of the PASCAL VOC detection challenge of 2007. The reason for using the data from 2007 is that it was the last time that testing annotations were provided. We need such annotations to enrich them with the different scenarios we have mentioned (Fig. 1). After doing such enrichment for training and testing data, we obtain the numbers of training windows and testing images per scenario summarized in Tab. 1.

In order to answer **Q2** we run experiments analogous to the scenario-based ones, but without actually distinguishing the scenario. In other words, we perform the type of experiments that PASCAL VOC challenge participants do, for the cases of RGB and OPP color spaces. Additionally, we not only present the

overall result on the full testing dataset but also the results of applying the overall classifiers (*i.e.*, the ones trained without taking into account the scenarios) to each considered scenario separately.

It is worth to mention that Felzenszwalb *et al.* method computes the HOG over the *max-gradient* as we have seen in Sect. 2, however, we compute separate HOG features for each OPP channel. Thus, our features are of a dimension three times higher than the usually used for HOG computation. Nevertheless, for a fair comparison we did similar experiments using the separate R, G and B channels in an analogous use to the three OPP channels. The results were basically analogous to the use of *max-gradient* for RGB, thus, the conclusions of this paper do not change. Accordingly, here on we will only focus on the usual procedure found in the literature, *i.e.*, computing the *max-gradient* for RGB. Note that while RGB channels are highly correlated ones, OPP ones are not.

For the training of any classifier we apply the bootstrapping method to collect hard negatives. We follow the scheme provided by the publicly available software of Felzenszwalb *et al.*, which collects all possible hard negatives until filling 3GB of working memory. In practice, this means to perform about ten bootstrappings.

In order to evaluate the obtained results, we follow the PASCAL VOC 2007 protocol, which is based on *precision-recall* (PR) curves and the associated *average precision* (AP). Please, refer to [6] for more details about such protocol.

In summary, the experiments to be done are:

- *Indoor, countryside* and *urban* classifiers: they are learnt from indoor images and applied to indoor images. The same for countryside and urban ones.
- *Overall* classifier: it is learnt from all the images but tested in different ways: on all the test images; only in the test images classified as *indoor*; only *countryside*; only *urban*.

These experiments must be run for OPP and RGB color spaces. Thus, we get a total of 14 PR curves and corresponding APs. Figure 2 shows all the PR curves in a meaningful way and Tab. 2 presents the corresponding APs. Additionally, we also applied each scenario-specifically-trained classifier to the other scenarios (not trained). We do not plot the corresponding PR curves for the sake of simplicity but we include the respective APs in Tab. 2.

4 Discussion

Results summarized in Fig. 2 and Tab. 2 allow to answer the questions **Q1** and **Q2** stated in Sect. 3.

Table 2 shows that AP in indoor scenarios is 1.7 points higher for OPP than for RGB when using only such type of scenarios for training. In the case of countryside the difference is even higher, 3.1 points. However, in urban scenarios RGB performs 0.6 points better.

A closer look to the PR curves (Fig. 2) for indoor, urban and countryside scenarios gives more detailed insight. In the case of indoor scenarios we appreciate that for the specifically trained and tested classifiers the difference between

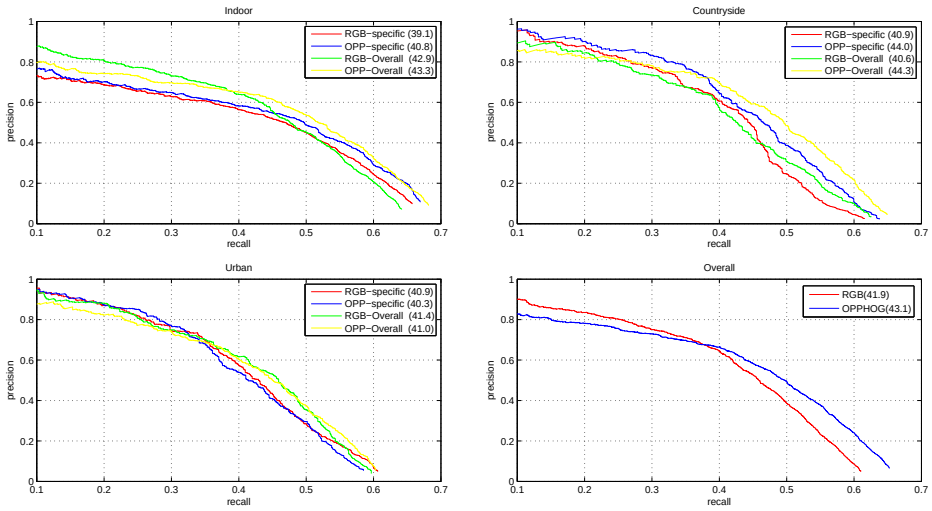

Fig. 2. Precision-recall (PR) curves obtained from the different experiments are shown: using RGB and OPP color spaces, for the indoor, countryside, urban and overall classifiers. The average precision (AP) of each PR curve is the number shown inside the respective parenthesis. The PRs of the specific classifiers are plotted together with the PRs of the overall classifiers applied only in the corresponding specific scenarios.

	RGB			OPP		
	Indoor	Countryside	Urban	Indoor	Countryside	Urban
Indoor	39.1	21.8	21.2	**40.8**	23.4	22.8
Countryside	22.0	40.9	31.1	24.9	**44.0**	33.4
Urban	29.9	34.9	**40.9**	33.3	39.8	40.3
Overall	41.9			**43.1**		
	42.9	40.6	**41.4**	**43.3**	**44.3**	41.0

Table 2. Average precision (AP) in % of the different trained and tested classifiers. Indoor/Countryside/Urban/Overall in the first column refer to the training step, while Indoor/Countryside/Urban in the second row refer to testing. Bold numbers indicate the higher APs comparing the counterpart RGB and OPP results. For the overall classifiers we not only include the overall APs, but also the APs corresponding to apply such classifiers only to specific scenarios during testing.

OPP and RGB is higher for higher recall. This fact is not captured by the AP computation method used in PASCAL VOC 2007 detection challenge. Note, that detection systems are usually interested in having higher recall. In countryside scenarios we observe an analogous situation, but with higher differences. In the case of urban scenarios we see that the specifically trained classifiers are pretty similar along the whole PR plot.

From these observations we conclude that the answer to question **Q1** is: *for indoor and countryside scenarios OPP color space performs better than RGB,*

while for urban scenarios it seems that there is not a clear preference for mid-to-high recalls. The major benefit of OPP is for countryside scenarios. Interestingly, OPP color space is the result of human evolution inside primitive indoor and countryside environments, not urban ones, where humans were targets of interest among others. Primitive indoor scenarios are of different background than modern ones. However, countryside colors remain constant. Of course, we don't argue here that our experiments are supporting psychological/evolutive claims about the human vision system, we only want to point out here what in our modest opinion is an interesting fact.

Regarding question **Q2**, Tab. 2 shows that when jointly using all human windows and backgrounds for training, the AP is 1.2 points higher for OPP than for RGB. Again, by a closer look to PR curves (Fig. 2) for the overall case, we observe that the major benefit of OPP comes for recalls over 40%, *e.g.*, for a recall of the 50% we obtain about ten points more of precision with OPP. We can also assess the performance of these overall classifiers focused on our specific scenarios. We observe then that for the indoor ones, for recalls below the 40% RGB is giving higher precision, however, over such recall the situation changes. The AP is 0.4 points higher for OPP than for RGB. The case of countryside scenarios is analogous but here the OPP starts to offer better precision before, approximately for recalls higher than the 22%. The AP is 3.7 points higher for OPP. In urban scenarios precision is higher with RGB than with OPP for recalls lower than approximately the 30%, however, over such recall OPP and RGB behave pretty similar. The AP is 0.4 points higher for RGB.

From these observations we conclude that the answer to question **Q2** is: *combining data coming from different scenarios during training helps to potentially obtain benefits from OPP over RGB, however, the final benefits will only be obtained if the classifier is used in indoor and countryside scenarios*. Note that the best scenario for OPP, i.e., countryside according to our experiments, is the less represented in the training of overall classifiers (Tab. 1). During testing, countryside and urban scenarios are, basically, equally represented, but indoor scenarios gain in testing presence (Tab. 1), which probably is the reason for OPP offering an overall improvement over RGB (countryside cases help AP for OPP while urban cases help RGB).

In summary, using OPP for human detection is worth out of urban scenarios, specially for countryside. Examining Tab. 2 one could be also tempted to conclude that overall detectors outperform the specifically trained ones, however, we think that this can be only an effect of the number of examples and counter-examples during training. What is clear (and expected), however, is that classifiers trained only in one type of scenario perform poorly in the other types of scenarios.

5 Conclusion

In this paper we have investigated the effect of using the opponent color space, which is based on the human vision system, for person detection. We have taken

as baseline person detector the HOG and part-based method proposed by Felzenszwalb *et al.*. Then, by following the protocols of the PASCAL VOC challenge of 2007, applied to the *person* class, we have collected experimental results that state that opponent color space is a better choice for computing HOG in indoor and, specially, countryside environments. In urban scenarios, there is no clear benefit. Interestingly, indoor and countryside scenarios, those in which the human visual system was *designed* by evolution. The combination of opponent color scape and Felzenszwalb *et al.* method as well as the scenario-based study are new up to the best of our knowledge.

As future work we plan to combine scenario-specific trained classifiers with image classifiers so that, given a new image of unknown type, we can compute the type of scenario to which it belongs (or a probability for each type) and apply a selection methodology (or a fusion scheme) in order to obtain the best benefit of the learned classifiers.

Acknowledgments This work has been supported by the Spanish Government projects TRA 2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018).

References

1. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with dscriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1627–1645
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA (2005)
4. van de Sande, K., Gevers, T., C.M. Snoek: Evaluating color descriptors for object and scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1582–1596
5. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision, Kerkyra, Greece (1999)
6. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. Journal on Computer Vision **88**(2) (2010) 303–338
7. Fischler, M., Ekschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers**100**(22) (1973) 67–92
8. Krauskopf, J., D.R. Williams, D.W. Heeley: Cardinal directions of color space. Vision Research **22**(9) (1982) 1123–1132
9. J.D. Mollon: "tho' she kneel'd in that place where they grew ..." the uses and origins of primate colour vision. Journal of Experimental Biology **146**(1) (1989) 21–38