Opponent Colors for Human Detection

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. Human detection is a key component in fields such as advanced driving assistance and video surveillance. However, even detecting non-occluded standing humans remains a challenge of intensive research. Finding good features to build human models for further detection is probably one of the most important issues to face. Currently, shape, texture and motion features have deserve extensive attention in the literature. However, color-based features, which are important in other domains (*e.g.*, image categorization), have received much less attention. In fact, the use of RGB color space has become a kind of choice *by default*. The focus has been put in developing first and second order features on top of RGB space (*e.g.*, HOG and co-occurrence matrices, resp.). In this paper we evaluate the opponent colors (OPP) space as a biologically inspired alternative for human detection. In particular, by feeding OPP space in the baseline framework of Dalal *et al.* for human detection (based on RGB, HOG and linear SVM), we will obtain better detection performance than by using RGB space. This is a relevant result since, up to the best of our knowledge, OPP space has not been previously used for human detection. This suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., also on top of OPP space, *i.e.*, as we have done with HOG in this paper.

1 Introduction

Human detection is a key component in fields such as advanced driving assistance [1–3] and video surveillance [4–6]. Detecting humans in images is quite challenging because of their intra-class variability, the diversity of backgrounds and the different image acquisition conditions. Even detecting non-occluded humans that are standing, is still a hot topic of research. In order to improve human detection results we can focus on *classification*, *i.e.*, on building a classifier that given an image window decides if it contains a human or not. Nowadays, most successful classification processes for human detection follow the learning-from-examples paradigm [1, 2]. For instance, Dalal *et al.* [7] proposed a holistic classifier that relies on histograms of oriented gradients (HOG) as features and linear support vector machines (linear SVM) as learning algorithm, which still remains as a competitive baseline method for comparison with new human classifiers [2, 8].

Finding good features for developing a human classifier is a major key for its success. Focusing on human appearance, different sets of features try to exploit

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 363–370, 2011.

(combinations of) cues such of shape and texture [4]. However, although color information deserves special attention in domains such as segmentation and category recognition [9, 10], it has not been explored in deep for human detection. In fact, the baseline classifier of Dalal *et al.* [7] uses standard RGB. In particular, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. Dalal *et al.* reported that similar results were obtained using LAB space. This approach has been the common way of using color for human detection since then [4, 11–15] and, as a matter of fact, it has been considered as pretty similar to the use of the image intensity in cases where color information was not available [2].

Human beings do not rely on long (L), middle (M) and short (S) wavelength channels (RGB-like) separately for color perception. In order to increase subsistence, evolution provided the human retina with ganglion cells that combine L, M and S channels to work in *opponent-colors-space* mode for enhancing the visual detection of events of interest as well as compressing the color information of L, M and S *acquisition cells* [16–18]. Such compressed color information is sent through the optical nerve to the brain for later decompression and interpretation. Accordingly, in this paper we evaluate the opponent colors (OPP) space as a biologically inspired alternative for human detection. In particular, by feeding OPP space in the baseline framework of Dalal *et al.*, we will obtain better detection performance than by using RGB space. Besides, this finding is reinforced by the work in [10], where K. van de Sande *et al.* show that applying a scale invariant feature transform (SIFT [19]) to OPP space is the best *a priori* option in the context of image category recognition. Note, that HOG is a SIFT inspired descriptor.

For our current work, as Dalal *et al.*, we have used the so-called INRIA human dataset. This dataset contains color images and still is widely used for benchmarking. To support our claim we not only present so-called *per window* evaluation on INRIA human dataset, but also *per image* evaluation as highly recommended in [8].

We argue that altogether is a relevant result since, up to the best of our knowledge, OPP space was not previously used for human detection. Thus, with the aim of enriching feature space for human classifiers, our work suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., on top of OPP space, *i.e.*, as we have done here with HOG.

The rest of the paper is organized as follows. In section 2 we define the OPP space. In section 3 we summarize the details of the human detector developed for our experiments. In section 4 we draw the experiments and discuss the corresponding results. Finally, section 5 summarizes the main conclusions.

2 Opponent Colors Space

In the late 19th century, E. Hering noted that the four hues red, green, yellow and blue are fundamental in the sense that they cannot be described as mixtures

of other hues. Then, he stated that there were three types of photo receptors: white-black, yellow-blue and red-green [16]. Nowadays we know that there are not such *image acquisition cells* in human vision. However, Hering was right in postulating the *computation of opponent colors* (*i.e.*, red *vs* green and yellow *vs* blue) in human color vision.

Contemporary science of human vision states that color photo receptors at the retina (*i.e.*, cones) are sensitive to long (L-cone), middle (M-cone) and short (S-cone) wavelengths. A single cone is color blind since its activation depends on both the wavelengths and intensity of the stimulus. A comparison of the signals from different classes of photo receptors is therefore the most basic computational requirement of a color vision system. The existence of cone-opponent retinal ganglion cells that perform such comparisons is well established for human vision.

In particular, 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 [17]. 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 [18].

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 [10]:

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

for R, G and B running on values in $[0, 1]$.

3 Human Detector

A *human detector* is composed of a *human classifier* learnt from a training set by using specific *features* and a *learning machine*. With this classifier we *scan a given image* looking for humans. Since multiple detections can be produced by a single human, we also need a mechanism to *select the best detection*. The procedures we use for feature extraction, machine learning, scanning the images, as well as selecting the best detection from a cluster of them, are briefly reviewed in this section.

Human classifier. We follow the settings suggested by Dalal *et al.* for computing HOG features and learning the human classifier using a linear SVM. Such approach remains competitive [2, 8] and, in fact, is the core from which many

new proposals are developed [4, 14]. However, Dalal *et al.* as well as in many following works [4, 11–15], compute HOG on top of RGB space. More specifically, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. We argue that the *max* operation basically is throwing away the color information, *i.e.*, only some sort of luminance contrast is captured by HOG. Accordingly, we propose to replace the features considered by Dalal *et al.* so that color information is also captured.

Our proposal is twofold. First, we remove the *max* operation, *i.e.*, HOG are applied to each color channel separately and, then, the corresponding feature vectors are concatenated to form a single feature vector. Such three-channels HOG are then the input that the linear SVM will use to learn the human classifier. Second, we propose the use of OPP space instead of RGB one. We will see that both ideas are essential to improve human classification performance.

Image scanning. In order to perform multi-scale human detection we use the extended *pyramidal sliding window* strategy as proposed in Dalal's PhD [20]. The original image is scaled by a factor s^i to obtain the image corresponding to the pyramid level i. Then, given a pyramid level, we must shift the search window along the horizontal and vertical directions with a given stride $\Delta = (\delta_x, \delta_y)$ pixels. The smaller the s and Δ parameters, the finer the sliding window search. Using a finer search we can expect better detection performance. However, this is to the expense of a higher processing time. Dalal set $s = 1.2$ and $\Delta = (8, 8)$. In our work we found $s = 1.05$ and $\Delta = (4, 4)$ pixels a better tradeoff between processing time and detection performance.

Select the best detection. In multi-scale human detection a single person can be detected several times at slightly different positions and scales. Since a unique detection per human is desired, multiple overlapped detections should be grouped by a clustering or *non-maximum-suppression* procedure. In this case, we don't follow the Dalal's proposal in [20]. Instead, we rely on the iterative confidence- and overlapping clustering approach of Laptev [21], which is a simpler and faster technique than Dalal's proposal and yields similar results.

4 Experiments

4.1 Human Dataset

We rely on the widely used INRIA person dataset of color images for our experiments. This dataset shows a wide range of human variations in pose, clothing, occlusions as well as complex backgrounds. Moreover, the dataset is divided in separated sets of null intersection for training and testing.

The training set contains 2,416 *positive* samples consisting in image windows (original and vertical mirror), each one containing a person framed by certain

Fig. 1. Positive (humans) and negative (background) windows from INRIA dataset

amount of background. Positives are of the same size (*canonical detection window*), although many of them come from an isotropic down scaling. We term this set of windows as $\mathcal{W}_+^{\text{train}}$. For collecting *negative* samples, *i.e.*, image windows that do not contain persons, there are available 1,218 human-free images. We term this set of images as $\mathcal{I}_-^{\text{train}}$. The testing set consists of: (1) $\mathcal{I}_-^{\text{test}}$: 453 human-free images; (2) $\mathcal{I}_+^{\text{test}}$: 288 images containing labelled persons (ground truth); (3) $\mathcal{W}_+^{\text{test}}$: 1,126 positives analogous to the ones in $\mathcal{W}_+^{\text{train}}$ after cropping and mirroring the ground truth of $\mathcal{I}_+^{\text{test}}$.

4.2 Training

We use the standard training procedure for the INRIA dataset [7, 20]. First, we collect random negative windows from the images in $\mathcal{I}_-^{\text{train}}$ (10 windows per image to have 12,180 negatives) and down scale them to the size of the canonical detection window; let's call this set of windows $\mathcal{W}_-^{\text{train}}$. Then, given the sets $\mathcal{W}_+^{\text{train}}$ and $\mathcal{W}_-^{\text{train}}$, we compute the HOG of such labelled windows on top of the desired color space, and learn the human classifier using the linear SVM. Finally, we run the corresponding human detector on $\mathcal{I}_-^{\text{train}}$ in order to follow the recommended *bootstrapping* technique, *i.e.*, to append the set $\mathcal{W}_-^{\text{train}}$ with *hard negative windows* and re-train the human classifier. We apply two bootstrapping iterations. Figure 1 shows positive and negative training samples.

4.3 Evaluation

In our experiments we use two widely extended methods of evaluation: *per window* and *per image*. In per window evaluation we asses the results of the human classifier when applied to the $\mathcal{W}_+^{\text{test}}$ and the images in $\mathcal{I}_-^{\text{test}}$. Let $\text{P}^{\#}$ be the cardinality of $\mathcal{W}_+^{\text{test}}$, and let's term as P^{TP} the number of elements in $\mathcal{W}_+^{\text{test}}$ classified as *humans* (*i.e.*, total of so-called true positives). Let $\text{N}^{\#}$ be the total number of windows processed by applying the pyramidal sliding window technique to the images in $\mathcal{I}_-^{\text{test}}$ (for each image more than one million of windows are usually processed), and let's term as N^{FP} the number of such windows classified as *humans* (*i.e.*, total of so-called false positives). Then, we define the per window detection rate as $\text{DR}^{\text{pw}} = \text{P}^{\text{TP}}/\text{P}^{\#}$, $\text{DR}^{\text{pw}} \in [0, 1]$. Corresponding miss rate is defined as $\text{MR}^{\text{pw}} = 1 - \text{DR}^{\text{pw}}$. Analogously, we define the false positives per window as $\text{FP}^{\text{pw}} = \text{N}^{\text{FP}}/\text{N}^{\#}$, $\text{FP}^{\text{pw}} \in [0, 1]$. We remark that for any given image window, the human classifier returns a real value that we threshold with a fixed value t in order to classify the window as of type *human* or *non-human*.

Thus, $\mathrm{DR}^{\mathrm{pw}}$ and $\mathrm{FP}^{\mathrm{pw}}$ are functions of t. This allows to plot evaluation curves $\mathrm{E}^{\mathrm{pw}}(t) = (\mathrm{FP}^{\mathrm{pw}}(t), \mathrm{MR}^{\mathrm{pw}}(t))$ (so-called ROCs) that show the tradeoff between the miss rate and the false positives per window for each t.

However, some researchers show that it may be more realistic to follow per image evaluation [8]. In this case, not only the human classifier is evaluated but the whole human detector. In particular, the sets $\mathcal{I}_+^{\mathrm{test}}$ and $\mathcal{I}_-^{\mathrm{test}}$ are seen as a single set of images, $\mathcal{I}^{\mathrm{test}}$, where the human detector is run. Then, the set of detections is compared with the ground truth for counting how many of such detections are true positives (T^{TP}) and how many are false positives (T^{FP}). If $\mathrm{I}^{\#}$ is the cardinality of $\mathcal{I}^{\mathrm{test}}$ and $\mathrm{H}^{\#}$ the number of labelled humans in $\mathcal{I}_+^{\mathrm{test}}$, then we can define the per image detection rate as $\mathrm{DR}^{\mathrm{pi}} = \mathrm{T}^{\mathrm{TP}}/\mathrm{H}^{\#}$ ($\mathrm{DR}^{\mathrm{pi}} \in [0, 1]$; per image miss rate $\mathrm{MR}^{\mathrm{pi}} = 1 - \mathrm{DR}^{\mathrm{pi}}$) and the false positives per image as $\mathrm{FP}^{\mathrm{pi}} = \mathrm{T}^{\mathrm{FP}}/\mathrm{I}^{\#}$. In order to determine if a detection overlaps sufficiently with a labelled human of $\mathcal{I}_+^{\mathrm{test}}$ we follow the so-called PASCAL criteria [8] (also for bootstrapping during training). Now, analogously to $\mathrm{E}^{\mathrm{pw}}(t)$ we can define the evaluation curve $\mathrm{E}^{\mathrm{pi}}(t) = (\mathrm{FP}^{\mathrm{pi}}(t), \mathrm{MR}^{\mathrm{pi}}(t))$; $\mathrm{FP}^{\mathrm{pi}}(t)$ can be greater than one.

4.4 Devised Experiments

We train four types of classifiers: RGB-*max*; RGB-*3ch*; OPP-*max* and OPP-*3ch*. The OPP *vs* RGB refers to the used color space. The *3ch* stands for computing HOG for each color channel separately and then concatenate the three feature vectors into a single one. The *max* stands for computing HOG by taking into account, at each pixel, only the gradient of highest magnitude among the color channels, *i.e.*, the usual approach introduced by Dalal *et al.*

Since collecting negatives during training involves a random selection, obtained classifiers can vary from train to train. Therefore, for each type of classifier we repeat the training and further evaluation five times. This gives five curves per classifier (20 curves), thus, we condense the results for each classifier in the respective mean ± standard deviation curves for both per window ($\mathrm{E}^{\mathrm{pw}}(t)$) and per image ($\mathrm{E}^{\mathrm{pi}}(t)$) evaluation. Figure 2 summarizes the obtained results.

4.5 Discussion

We point out two main observations: (1) OPP-*3ch* clearly outperforms RGB-*3ch/max*; (2) the *max* operation throws away the color information. Let us argue these observations. Per image and per window evaluation show that OPP-*3ch* outperforms RGB-*3ch/max*, especially at the usual points of interest, *i.e.*, $\mathrm{FP}^{\mathrm{pw}} = 10^{-4}$ and $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$. At $\mathrm{FP}^{\mathrm{pw}} = 10^{-4}$ OPP-*3ch* has an average miss rate of 0.0358, while for RGB-*3ch* is 0.0806 and for RGB-*max* 0.1111. At $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$ OPP-*3ch* has an average miss rate of 0.1391, while for RGB-*3ch* is 0.1981 and for RGB-*max* 0.1889. Moreover, the *max* operation removes the difference between RGB and OPP spaces. Besides, per image evaluation shows

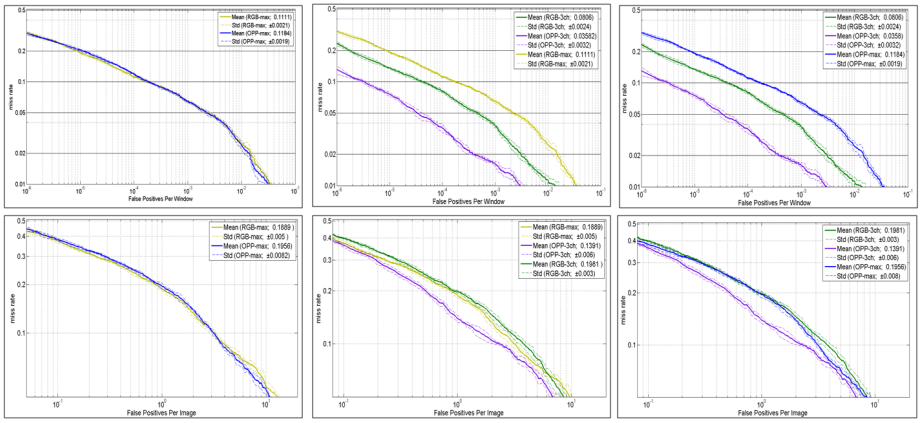

Fig. 2. Per window (top) and per image (bottom) evaluation using logarithmic scales. Values at usual points of interest are included, *i.e.*, 10^{-4} FPPW and 10^{0} FPPI, resp.

that both the *3ch* and the *max* configurations are similar for the RGB case, but quite different for OPP, where *3ch* clearly wins. For instance, the average miss rate of OPP-*3ch* at $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$ is 0.1391 while for OPP-*max* it is 0.1956.

5 Conclusions

In this paper we have explored the use of the biologically inspired opponent color space as the basis to obtain better features for human detection. In particular, we have seen that by feeding such a color space in the HOG+LinearSVM baseline classifier, we obtain better results than by following the common practice of using RGB color space. This conclusion is based on per window and per image evaluation over the widely used INRIA dataset. We think that this is a relevant finding, because, up to the best of our knowledge, opponent color was not previously used for human detection. Moreover, co-occurrence matrices, self-similarity features, etc., could be computed in the future on top of opponent color space in order to further improve human detection results.

Acknowledgments. This work was supported by the Spanish Government (projects TRA2007-62526/AUT, TRA2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018)).

References

1. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(7), 1239–1258 (2010)
2. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: survey and experiments. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12), 2179–2195 (2009)

3. Gandhi, T., Trivedi, M.M.: Pedestrian protection systems: issues, survey, and challenges. IEEE Trans. on Intelligence Transportation Systems 8(3), 413–430 (2007)
4. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: Int. Conf. on Computer Vision, Kyoto, Japan (2009)
5. Jones, M., Snow, D.: Pedestrian detection using boosted features over many frames. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
6. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Int. Journal on Computer Vision 63(2), 153–161 (2005)
7. 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)
8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: IEEE Conf. on Computer Vision and Pattern Recognition, Miami Beach, FL, USA (2009)
9. Álvarez, J.M., Gevers, T., López, A.M.: Learning photometric invariance for object detection. Int. Journal on Computer Vision 90(1), 45–61 (2008)
10. van de Sande, K., Gevers, T., Snoek, C.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
11. Oliveira, L., Nunes, U., Peixoto, P.: On exploration of classifier ensemble synergism in pedestrian detection. IEEE Trans. on Intelligence Transportation Systems 11(1), 16–27 (2010)
12. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: Int. Conf. on Computer Vision, Kyoto, Japan (2009)
13. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, Miami Beach, FL, USA (2009)
14. 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)
15. Leibe, B., Cornelis, N., Cornelis, K., Gool, L.V.: Dynamic 3D scene analysis from a moving vehicle. In: IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, USA (2007)
16. Hering, E.: Outlines of a theory of the light sense (translated by L.M. Hurvich and D. Jameson). Harvard University Press, Cambridge (1964)
17. Krauskopf, J., Williams, D.R., Heeley, D.W.: Cardinal directions of color space. Vision Research 22(9), 1123–1132 (1982)
18. Mollon, J.D.: "tho' she kneel'd in that place where they grew... " the uses and origins of primate colour vision. Journal of Experimental Biology 146(1), 21–38 (1989)
19. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision, Kerkyra, Greece (1999)
20. Dalal, N.: Finding people in images and videos. PhD Thesis, Institut National Polytechnique de Grenoble / INRIA Rhône-Alpes (2006)
21. Laptev, I.: Improving object detection with boosted histograms. Image and Vision Computing 27(5), 535–544 (2009)