Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Leveraging colour segmentation for upper-body detection

Stefan Duffner[*], Jean-Marc Odobez

Idiap Research Institute, 1920 Martigny, Switzerland

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form
25 September 2013
Accepted 27 December 2013

Keywords:
Detection
Colour segmentation
Upper-body

ABSTRACT

This paper presents an upper-body detection algorithm that extends classical shape-based detectors through the use of additional semantic colour segmentation cues. More precisely, candidate upper-body image patches produced by a base detector are soft-segmented using a multi-class probabilistic colour segmentation algorithm that leverages spatial as well as colour prior distributions for different semantic object regions (skin, hair, clothing, background). These multi-class soft segmentation maps are then classified as true or false upper-bodies. By further fusing the score of this latter classifier with the base detection score, the method shows a performance improvement on three different public datasets and using two different upper-body base detectors, demonstrating the complementarity of the contextual semantic colour segmentation and the base detector.

1. Introduction

The automatic detection and localisation of humans in digital images has been of increasing interest in the past decades. Applications include video-surveillance, human–computer interaction, image indexation and retrieval, and advanced driver assistance systems. According to the scenario of interest as well as the image resolution and quality, different body parts may be more visible and thus more detectable than others. In this regard, localising people has mainly been achieved by building detectors for faces, upper bodies, full bodies, or combination of these. Currently, *face detection* algorithms produce relatively few false positives for near-frontal head poses, but suffer from degraded performance under natural and arbitrary poses that people take when their attention is not directed towards the camera, and are not suitable for back views. In other scenarios like video-surveillance, *full-body* detectors are more appropriate, and state-of-the-art methods can cope relatively well with most of the common challenges such as articulated body poses, low image resolution, or poor lighting conditions. Still, *upper-body* detectors are of particular interest when there are frequent occlusions, like in crowded scenes, or in environments where the lower part of the body is not visible, like in TV broadcasts and movies or in video-conferencing applications (see Fig. 1). In this paper, we are interested in the latter type of scenarios and therefore we focus on upper-body detection, although the proposed method could be applied to face or body detection as well.

Motivations: The main idea of the proposed approach is to use soft colour-based semantic segmentation maps to distinguish *true* positives from *false* positives coming from a pre-trained upper-body shape detector based on Histograms of Oriented Gradients (HOG). In fact, although such HOG detectors have proven to be generic and efficient, their application on real-world data still results in false positives like the ones depicted in Fig. 2, which obviously contain upper-body like shape information. One can expect that a colour segmentation of these detected patches would be quite different than those obtained on trained upper-body patches. On the contrary, the segmentation of missed detection (examples are given in Fig. 2) would exhibit a closer match.

In spite of this intuition, colour has not been used frequently in practice. There are several reasons for that. A major one is that the colour values are often not a discriminative factor. For instance skin colour might be discriminative, but not the clothing ones. An alternative view is that colour information lies more in the spatial segmentation they induce rather than in the colour values themselves. However, exploiting plain image colour segmentation maps to define object colour features is difficult, as the number of regions within an object region is highly variable and usually depends on some threshold and on the highly variable colour contrast *within* object segments and between the object and the background.

To address the above issues, we propose to exploit the results of a probabilistic multi-class colour segmentation algorithm applied to candidate object patches generated by a first (*e. g.* shape-based) object detector algorithm. The benefits are as follows. Using a fixed number of classes with semantic meaning (for upper-body, face, hair, clothing and background), we avoid having to handle unpredictable segmentations in regions like background or clothing that may introduce more noise than discriminative information. In addition, working on candidate detections allows the colour

[*] Corresponding author at: Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France.

E-mail addresses: stefan.duffner@liris.cnrs.fr (S. Duffner), jean-marc.odobez@idiap.ch (J.-M. Odobez).

0031-3203/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.12.014

Fig. 1. Example scenarios for upper-body detection in images: video-conferencing and entertainment (left), movies (right). Face detectors and full-body person detectors would have difficulties due to occlusions and difficult head poses.

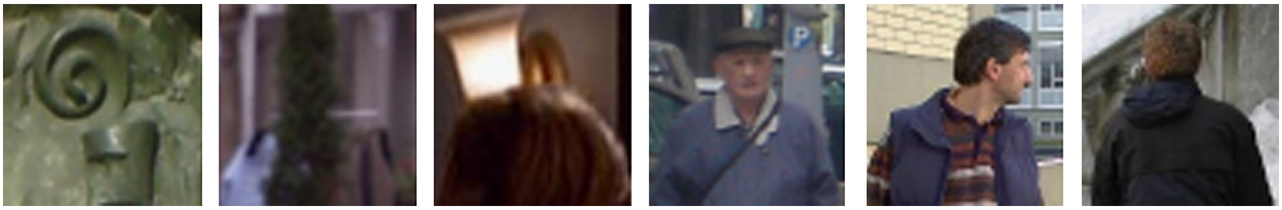

Fig. 2. Typical false positives and missed detections of the state-of-the-art HOG-based upper-body detector from the Calvin project (Eichner et al. [1] relying on Felzenszwalb et al. [2]), at a precision of 90%.

segmentation to be conducted *in a particular context*. In particular, one can exploit spatial priors over the locations of class pixels, ultimately leading to better class colour models and finer segmentations. Furthermore, one additional advantage of the probabilistic approach we take is the definition of priors on class colour models, which brings information to the segmentation whenever appropriate, like a more informed prior for the skin/face part and a broader prior for the background. Finally, by combining the colour segmentation information from the above algorithm with the score from the first high-recall low-precision detector, we expect to retain the high recall property while improving on the precision. This will be shown in the results.

Related work: Numerous object detection methods have been proposed during the last years. One could classify them into two main categories: *holistic* and *part-based* approaches. Earlier holistic approaches relied on Haar wavelet related features [3,4] and a subsequent Support Vector Machine (SVM) classifier or cascaded classifiers based on Adaboost. Later, HOG-based visual features have shown to give better performance for visual object detection tasks, including person [5], and upper-body detection [6–9] as they are capturing the characteristic omega shape of a human upper-body. Moreover, combinations of different low-level visual feature types have been shown to improve the overall detection performance, for example HOG with Haar-like [8] or Local Binary Patterns (LBP) [10] features. More recently, part-based detectors, *e.g.* [11–13,2,1,14], have proven to give better results thanks to their higher ability to handle body parts alignment variabilities.

Colour information has seldom been incorporated into full or upper-body detectors. For example, Micilotta et al. [12] used temporal adaptive skin colour models to reduce the number of false detections when tracking upper-bodies in video sequences. Extensions of HOG features with colour information have been proposed for person detection [15,16]. However, by definition these approaches consider colour *locally*, that is, without integrating a wider spatial context as we propose. The closest work to ours is that of Ramanan [17]. They applied a graph-cut colour segmentation algorithm to image patches produced by an object detector (of faces, persons, cars), and then classified the resulting binary segmentation map using a linear SVM. Although relying on the same hypothesis-testing scheme as ours, this approach suffers from several limitations: it only performs a binary segmentation, it does not integrate prior colour information when available, and it does not fuse the detection score relying on the segmentation feature with the score of the initial detector. Our experiments show that these points are indeed important to improve the performance.

Contributions: We present an effective technique for improving the precision of a first-stage upper-body detector using colour segmentation information. It relies on a *multi-class* probabilistic colour segmentation algorithm that leverages the *context* provided by candidate detections using *spatial* as well as *colour* prior distributions for different *semantic* object regions. A thorough evaluation on three different data sets using two different upper-body detection algorithms demonstrates the validity of our approach.

2. Upper-body detection with colour segmentation

2.1. Overall approach

Fig. 3 outlines the proposed method which is explained in the caption. A key component is the segmentation algorithm which assigns to each pixel of a detection window a probability distribution over the four classes: skin, hair, clothing, and background. In the following, we first summarise the approach of Scheffler and Odobez [18] and its interest for the detection task. Then we describe how we exploited it and explain the segmentation representation that has been used, and finally, we present the different classification steps.

2.2. Bayesian segmentation algorithm

To conduct the segmentation of a given detected image, we use an approach derived from [18]. The advantage of this algorithm [18] is that it provides us with a semantic multi-class segmentation where both spatial and colour prior knowledge, learnt from training data, are integrated. It relies on the generative model of image colour pixels represented in Fig. 4. It is an extension of the Probabilistic Index Map (PIM) of [19]. However, a major difference is a full Bayesian treatment of the model by defining appropriate priors on colour palettes.[2] These priors are particularly useful for

[2] Another important difference was the inclusion of a Markov Random Field (MRF) regularisation term on the segmentation map (called Coherent PIM in [18]). However, we did not use it and do not present it here as it did not result in better detection rates and is computationally more demanding.

Please cite this article as: S. Duffner, J.-M. Odobez, Leveraging colour segmentation for upper-body detection, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2013.12.014

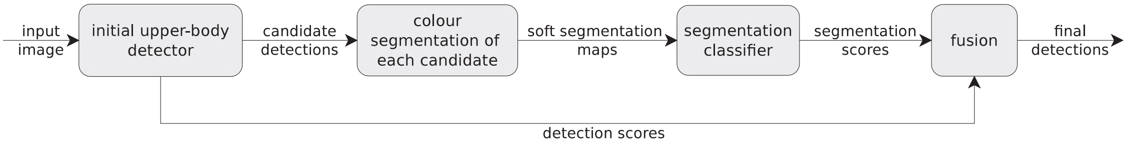

Fig. 3. Proposed scheme. Candidate upper-body image regions produced by an upper-body detector are colour segmented using a probabilistic algorithm. The resulting 4-class soft segmentation maps are classified by a linear SVM, whose score is further fused using a SVM with the detection score of the first stage detector to make a final decision.

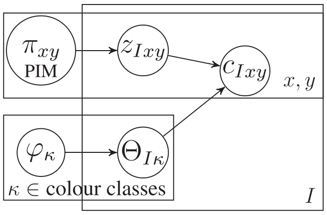

Fig. 4. Generative model of an image I (derived from [18]). Plate notation: a box indicates random variables which are repeated for each element of the index indicated at its bottom right. For each colour class κ, a specific palette (*i.e.* a colour distribution) for this image is drawn from a colour prior. The palettes are discrete distributions over colours (*i.e.* normalised histograms) parameterised by $\Theta_{I\kappa}$, and priors over $\Theta_{I\kappa}$ are Dirichlet distributions parameterised by φ_κ. Each pixel of coordinates x, y of the image I is assigned a colour class z_{Ixy} drawn from the Probabilistic Index Map (PIM) prior characterised by $p(z_{Ixy} = \kappa) = \pi_{xy}(\kappa)$, *i.e.* discrete distributions over classes at each pixel. Then, for each pixel, its actual colour c_{Ixy} in an image is generated from the palette corresponding to its class, *i.e.* $\Theta_{Iz_{Ixy}}$.

classes whose colours only occupy a small part of the colour space (like for skin), as they prevent the inference of an inappropriate class palette and ultimately the erroneous labelling of image pixels. This chromatic prior complements the class label PIM spatial prior and improves the segmentation results when there is more uncertainty in the spatial prior.

Model: we now more formally describe the main probabilistic terms included in the model. Fig. 4 and the corresponding caption explain the main elements of this generative model.

First, concerning the *colour* information, the colour priors, that are learnt beforehand, are Dirichlet distributions parameterised by φ_κ for each colour class κ (here skin, hair, clothing, and background). For a given image I and for each class κ, discrete colour distributions, called palettes (*i.e.* normalised histograms), parameterised by $\Theta_{I\kappa}$ are drawn according to

$$p_{\rm pal}(\Theta_{I\kappa}|\varphi_\kappa) = {\rm Dir}(\Theta_{I\kappa}|\varphi_\kappa). \tag{1}$$

Second, the *spatial* prior information is characterised by PIM prior maps defined by $p(z_{Ixy} = \kappa) = \pi_{xy}(\kappa)$ and the set of class labels, *i.e.* discrete distributions over classes at each pixel (see example in Fig. 5(b)). The class labels $\langle z_{Ixy}\rangle$ of a given image I are generated according to

$$p_{\rm PIM}(\langle z_{Ixy}\rangle|\langle\pi_{xy}\rangle) = \prod_{x,y}\pi_{xy}(z_{Ixy}). \tag{2}$$

Note that both the colour prior φ_κ and spatial PIM prior π_{xy} are image independent (in Fig. 4 they lie outside the plate indexed by I).

The class label z_{Ixy} at a pixel and the colour palette are then combined to draw the actual observed colour c_{Ixy}:

$$p_{\rm col}(c_{Ixy}|\Theta_{I\kappa}, z_{Ixy} = l) = \frac{1}{v}[\Theta_{Il}]_{{\rm bin}(c_{Ixy})}, \tag{3}$$

where v is a normalisation constant (the colour volume of histogram bin), $[\cdot]_a$ denotes the a-th component of a vector, and ${\rm bin}(\cdot)$ maps a colour vector to its bin in the colour histogram. As in [18] we use the RGB colour space and 16 histogram bins per channel. Fig. 5 (c) illustrates the inferred palettes for an example image.

Finally, in summary, the joint probability of the image random variables [4] $p(\langle z_{Ixy}\rangle, \langle\Theta_{I\kappa}\rangle, \langle c_{Ixy}\rangle|\langle\varphi_\kappa\rangle, \langle\pi_{xy}\rangle)$ is given by

$$p_{\rm PIM}(\langle z_{Ixy}\rangle|\langle\pi_{xy}\rangle)\prod_{\kappa} p_{\rm pal}(\Theta_{I\kappa}|\varphi_\kappa)\prod_{x,y} p_{\rm col}(c_{Ixy}|\Theta_{Iz_{Ixy}}). \tag{4}$$

Prior model training: In order to train the spatial PIM prior, the base detector has been applied to images that are not part of our evaluation datasets and for which images that are hand-segmented into semantic classes (here skin, hair, clothing, background) are available. Then, for each correct detection, the corresponding segmented image is cropped using the bounding box and resized to a common size (60 × 60). Finally, for each pixel of the 60 × 60 patch, the distribution π_{xy} of the class labels is computed using the set of cropped segmented images. Note that the individual images do not need to be fully annotated, as there are sometimes ambiguous pixels, *e.g.* near the segment boundaries.

For training the Dirichlet palette priors φ_κ for each class, we used the colour values of the annotated pixels from the spatial prior training mentioned above, as well as the Compaq skin database [20] containing pixel values for skin and non-skin colours.

Segmentation inference in a new image: the model is used to infer the actual colour palettes $\Theta_{I\kappa}$ and segmentation posterior probabilities $p(z_{Ixy}|\varphi_\kappa, \pi_{xy})$ of a given image, given the PIM spatial priors and the colour palette priors. Inference on a test image I is conducted by approximating the posterior beliefs over the class of each pixel z_{Ixy} and the palette of each class $\Theta_{I\kappa}$ using variational inference (details are given in [18]). This latter task is illustrated in Fig. 5. It depicts the input image, the spatial PIM prior, the inferred colour palettes and posterior segmentation maps and a visualisation of the maximum *a posteriori* segmentation (which is not used in our algorithm).

2.3. Upper-body segmentation representation

The algorithm described in the previous section has been applied to upper-body segmentation in the following way. We used four class labels: skin, hair, clothing, and background. We resorted here to discrete palettes for all labels rather than using continuous ones for the skin and hair classes as in [18]. We trained the class spatial priors from around 150 hand-segmented upper-body samples, which resulted in the PIM trained prior π^t_{xy}. This manual segmentation was done at the pixel level by two persons on individual images of a separate dataset. Ambiguous pixels (often located near segment boundaries) were not annotated, and annotation errors are smoothed out since π^t_{xy} is averaged over all segmented images. The Dirichlet palette priors for the skin, background and clothing[3] classes were learnt from the Compaq skin database [20] which contains more than 13 000 images, and from a hundred samples for the hair class.

The algorithm of [18] assumes to deal only with true positive. However, our task is to extract features that allow us to distinguish between “real” and “false” upper-body images. Therefore, we need to reduce the influence of the trained PIM prior π^t_{xy} during inference in order to avoid image patches not containing any

[3] The same prior was used for the clothing and background classes.

Please cite this article as: S. Duffner, J.-M. Odobez, Leveraging colour segmentation for upper-body detection, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2013.12.014

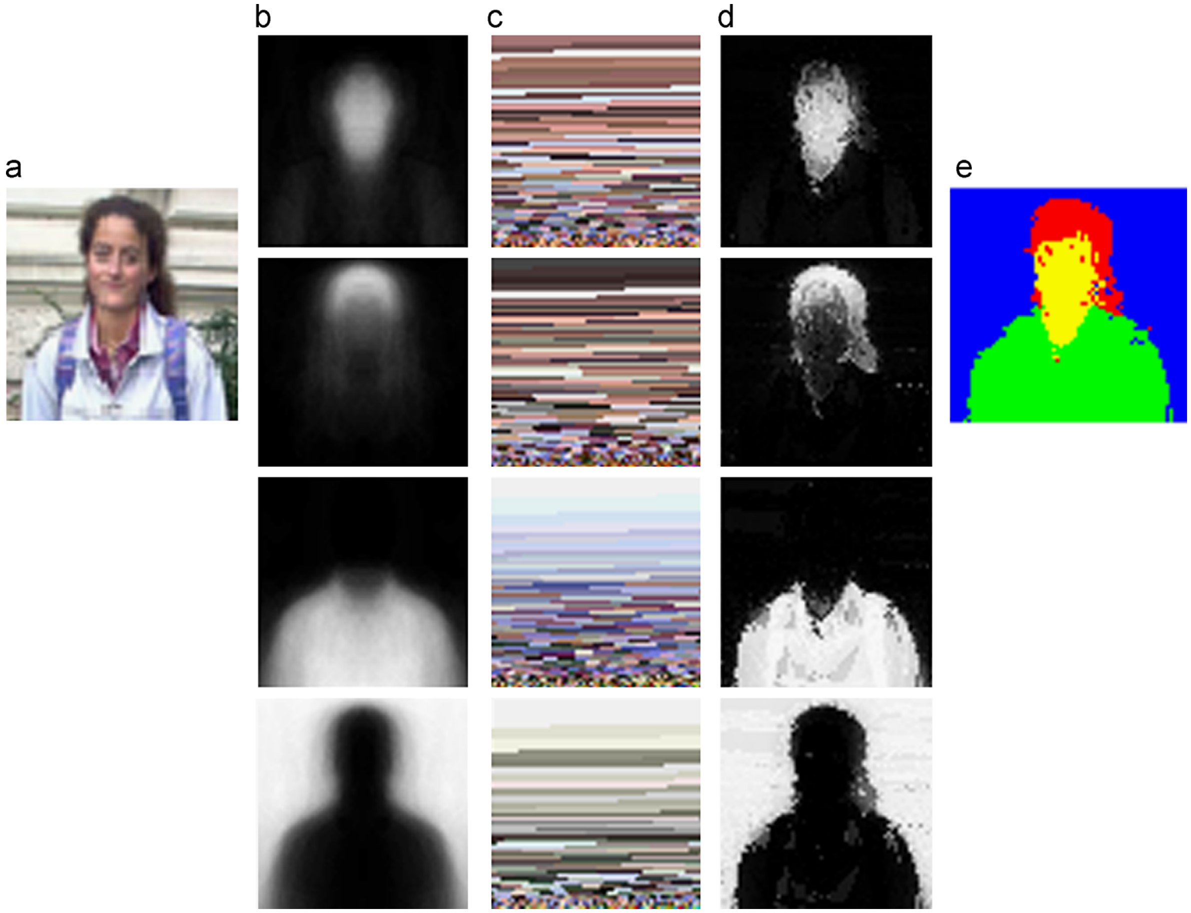

Fig. 5. An example illustrating the segmentation procedure: (a) input image; (b) spatial PIM prior (white: high probability, black: low probability); output of the segmentation inference, with (c) the inferred palettes (each colour bin is represented with a surface proportional to its probability) and (d) soft segmentation, *i.e.* for each class κ, the map over the (*x*,*y*) coordinate of $p(z_{Ixy}=\kappa|c_{Ixy},\pi_{xy},\varphi_\kappa)$, the posterior belief of the label κ given the observed image as well as the PIM and colour priors; (e) maximum *a posteriori* segmentation.

upper-body to be “forced” (by the PIM prior) to have an approximate upper-body shape. This was achieved by defining the actual PIM spatial prior π_{xy}, used when inferring soft segmentation maps on new test images, as a mixture of the uniform prior over the four classes:

$$\pi^u_{xy}(\kappa)=\frac{1}{4} \quad \forall\kappa \tag{5}$$

and of the trained prior π^t_{xy} according to

$$\pi_{xy}=\alpha\pi^u_{xy}+(1-\alpha)\pi^t_{xy}. \tag{6}$$

In practice we used $\alpha=0.2$. Such a weaker spatial prior does not alter significantly the segmentation of real upper-body images and makes the distinction between true and false positives easier.

Segmentation representation: The above algorithm produces soft-segmentation maps, *i.e.* the posterior probabilities at each pixel (x,y) and for each class κ: $p(z_{Ixy}=\kappa|\Theta_{I\kappa},\pi_{xy})$, as illustrated in Fig. 5d). Given an image patch we use these soft segmentations as the segmentation feature vector:

$$f_I=\langle f_{Ixy}\rangle \tag{7}$$

with

$$f_{Ixy}=p(z_{Ixy}|\Theta_{I\kappa},\pi_{xy}), \tag{8}$$

In practice, the size of the segmentation output (here 60 × 60) was normalised to a resolution of 30 by 30 pixels (by a simple resampling of a factor 2 without interpolation), giving a linearised vector f_I of $4\times30\times30=3600$ dimensions. Downsampling the segmentation output to a resolution of 30 by 30 pixels has only a small effect on the classification performance (as neighbouring points are highly correlated) and increases the execution speed.

2.4. Classification

Having the segmentation features of the image patches coming from the base upper-body detector, two classifiers need to be defined and trained: one that classifies the segmentation features, and one that fuses the resulting score with that of the base detector (*c.f.* Fig. 3). Here lies one of the key contributions of our paper, that is, to use these semantic colour segmentation features in order to improve on the performance of state-of-the-art upper-body detectors.

Colour segmentation classifier: This classifier receives the (3600-dimensional) soft-segmentation features f_I as input and is trained to distinguish between true and false upper-body images. Following previous works [5,17] and for efficiency reasons, we used a linear SVM as classifier. The signed distance to the separating hyper-plane of the trained SVM was interpreted as a score called “segmentation score” hereafter in the fusion process. Thus, the segmentation score s^s_I of an Image I is defined as

$$s^s_I=w_s\cdot f_I+b_s, \tag{9}$$

where w_s and b_s are the trained hyper-plane parameters.

The SVM was trained using automatically segmented image patches from around 8300 (5700 positive, 2600 negative) upper-body detections that have been manually labelled as true

Please cite this article as: S. Duffner, J.-M. Odobez, Leveraging colour segmentation for upper-body detection, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2013.12.014

upper-bodies or false detections. Detections have been labelled correct if there was sufficient overlap with the ground truth rectangle and incorrect otherwise. The image samples were obtained from web images and videos queried using Google, and they were different from the data used for evaluation.

Fusion classifier: The last step consists in fusing the detection score s_I^d provided by the base upper-body detector with the segmentation score s_I^s. An SVM trained on the concatenation of the two scores $s'_I = [s_I^d \; s_I^s]$ gives us the final score:

$$s_I = s'_I \cdot w + b, \qquad (10)$$

with w and b being the hyper-plane parameters of the trained SVM.

Here again, we used a linear SVM, trained using a dataset composed of samples from the same image and video material as above. This set contained around 2800 samples (1100 positive, 1700 negative) with scores from both detection and segmentation. With this approach, samples whose segmentations lead to a wrong classification but have a high detection score or vice versa (samples with an initially low detection score but with a good segmentation score) can still be classified correctly.

2.5. Computational speed

For simplicity and efficiency, we resampled the candidate upper-body images to a size of 60 by 60 pixels before segmentation. Upper-bodies that are smaller than that are up-sampled, but these cases are rare, and the segmentation algorithm still gives useful results. In terms of execution speed, the proposed algorithm requires around 10 ms to process a candidate upper-body image, *i. e.* to perform the segmentation and classification, on a 3 GHz 64 bit Intel processor. The Calvin detector runs at a speed of around 2.5 s per image (480 × 560). The Adaboost+HOG detector needs around 0.5 s per image (of the same resolution). Note that in the experiments, we only dealt with a few candidates per image (from 0 to 10 in general[4]).

3. Experimental results

3.1. Data

To demonstrate the effectiveness and validity of the proposed approach, we used three different public datasets containing images taken in different environments:

1. *InriaLite*[5], a subset of the INRIA person dataset[6] (the test set) containing 145 outdoor photographs of 219 persons in total, most of them entirely visible and viewed approximately from the front or from the back.
2. *TA2*[7], a set of 95 frames (containing 275 upper-bodies) extracted from the TA2 database, that is images from video-conference-like recordings of people sitting around a table (see Fig. 1, left).
3. *Web*[8], a set of 419 images, with 98 positive images containing 128 upper-bodies and 321 negative images. These images have been obtained from random queries to Google Images. They have been taken in all sorts of environments and conditions.

3.2. Algorithms

3.2.1. Base detectors

We used the following upper-body detectors to evaluate the impact of the base detector (see Fig. 3) on the usefulness of the proposed approach.

HOG-Adaboost: This is a holistic upper-body detector relying on HOG features and trained with Adaboost using the method of Laptev [21] in which Linear Discriminant Analysis is performed to train the weak classifiers. A detection cascade of 24 classifiers has been trained using around 1200 upper-body images from the INRIA person dataset (not contained in the InriaLite test set) and from Google Images, cropped to a resolution of 60 × 60 pixels and aligned using manual annotation.

Calvin: A more powerful detector based on deformable parts introduced by Felzenszwalb et al. [2] and applied to upper-body detection by Eichner et al. [1].

3.2.2. Compared methods

The following detection methods have been used for comparison:

No segmentation: In this method, only one of the base detectors mentioned (HOG-Adaboost or Calvin) is used, without exploiting any further colour segmentation.

Ramanan: We implemented as baseline the method of Ramanan [17]. Given the detection candidates produced by the base detector, it uses GrabCut, a graph-cut-based segmentation algorithm, to classify pixels into foreground and background, and a linear SVM to classify the resulting segmentations into *true* or *false* upper-body detection. We used the same spatial prior (for GrabCut, where we merged the skin, hair and clothing classes to define the foreground class) and the same SVM training parameters as for our proposed method. Table 1 outlines the major differences between this method and our approach.

Proposed approach: It refers to the method described in the previous section (*c.f.* Fig. 3), which combines the base detector (HOG-Adaboost or Calvin), the colour segmentation (Section 2), and the segmentation and fusion classifiers (Section 2.4).

3.3. Evaluation protocol

Performance measures: We evaluated the different algorithms using precision–recall curves. For the base detectors, these curves were obtained by varying the detection threshold, and following the protocol of the PASCAL Visual Object Classes Challenge [22]. More precisely, a detection window D is classified as correct if the condition $D \cap GT / D \cup GT > 0.5$ is satisfied, where GT denotes the ground-truth region. If there are several detections overlapping with the same person, only one is counted as correct, and the others are counted as false positives.

The threshold of the base detector has been set to a low value to obtain a high recall with a low precision. Then, the candidates images were processed by the corresponding method, and the curves were obtained by varying the threshold of the segmentation classifier (Ramanan method), or of the fusion classifier (proposed method).

Table 1
Differences between the proposed approach and the baseline (Ramanan [17]).

Ramanan	Proposed approach
2 region classes	4 region classes
Hard segmentation	*Soft* segmentation
Spatial smoothing in segmentation	*No* spatial smoothing
No classifier fusion	Fusion of classifier scores

[4] There was on average one or two times more detection than the true number of upper-bodies, *i.e.* the threshold of the base detector was set to have a precision between $\frac{1}{3}$ and 0.5

[5] http://www.robots.ox.ac.uk/vgg/software/UpperBody/.

[6] http://pascal.inrialpes.fr/data/human/.

[7] https://www.idiap.ch/dataset/ta2.

[8] Will be made publicly available.

Please cite this article as: S. Duffner, J.-M. Odobez, Leveraging colour segmentation for upper-body detection, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2013.12.014

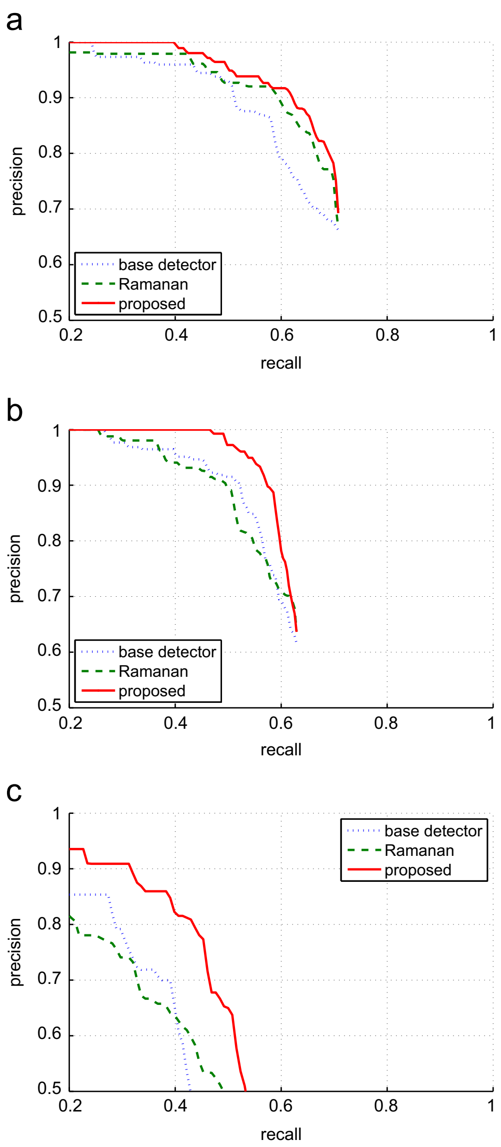

Fig. 6. Precision and recall of upper-body detection using the *Adaboost*+*HOG* base detector and different segmentation approaches. (a) Dataset 1 (InriaLite), (b) Dataset 2 (TA2), (c) Dataset 3 (Web).

Datasets: As mentioned in previous sections, the training data for the different algorithm steps was collected using different materials than the test datasets, which were only used for evaluating the algorithms.

3.4. Experiments

We have conducted two sets of experiments. In the first set, we compared the three approaches using the two base detectors (HOG+Adaboost and Calvin) on the three datasets. In the second one, we detail the individual performance of the different algorithmic steps (segmentation and fusion) on the results. Finally, we show the influence of the parameter α in the segmentation algorithm.

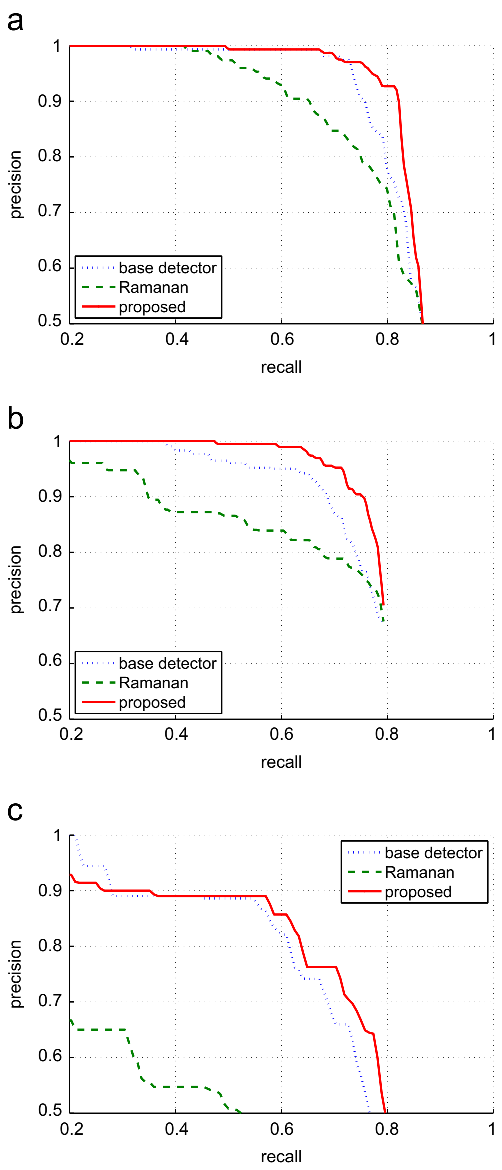

Fig. 7. Precision and recall of upper-body detection using the *Calvin* base detector and different segmentation approaches. (a) Dataset 1 (InriaLite), (b) Dataset 2 (TA2), (c) Dataset 3 (Web).

Table 2
Average precision (in %) of upper-body detection on the three different datasets using the different segmentation algorithms.

Detector	Method	Dataset			Relative increase
		1	2	3	
HOG+ Adaboost	No segmentation	65.79	58.44	40.66	
	Ramanan [17]	66.95	58.74	41.62	2.36
	Proposed approach	**68.19**	**61.26**	**48.63**	**19.60**
Calvin	No segmentation	83.10	75.75	70.38	
	Ramanan [17]	80.26	70.74	46.05	−34.57
	Proposed approach	**84.47**	**77.67**	**72.16**	**2.53**

Overall results: Figs. 6 and 7 show the resulting precision–recall curves for the Adaboost+HOG and Calvin detectors respectively. The proposed approach outperforms the baseline methods for both types of detectors and for all three datasets. Table 2 summarises these results. In many cases Ramanan's method even decreases the overall precision and recall. This might be due to the relatively high pose variability of upper-bodies, and the quite large difference in appearances across different datasets.

Overall, the relative increase of the mean average precision with the proposed approach is around 6%. Also, at a fixed precision of 80%, our proposed method increases the recall by around 10 percentage points for the Adaboost+HOG detector and by

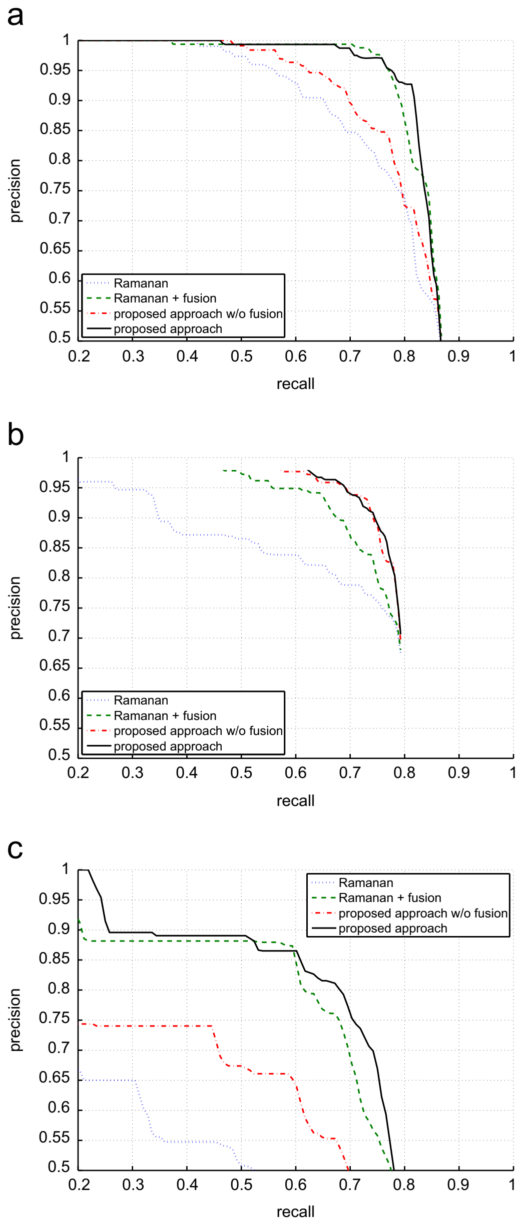

Fig. 8. Precision and recall of upper-body detection *with* and *without* the final classifier fusion (using the Calvin base detector). (a) Dataset 1 (InriaLite), (b) Dataset 2 (TA2), (c) Dataset 3 (Web).

around 5 percentage points for the Calvin detector. The smaller improvement observed on the InriaLite dataset with the Calvin detector could be due to the fact that this detector has been evaluated on this dataset, and thus the software that we obtained from the Web has been very probably trained on these data, resulting in some overfit. This gives a much smaller room for improvement using additional cues such as colour segmentation features.

Detailed results: In the second set of experiments, we applied the Calvin base detector and evaluated the performance of our approach *with* and *without* the final classifier fusion step. We compared it to the method of Ramanan [17] and extended his method with the same classifier fusion step as for our method. Fig. 8 shows the resulting precision–recall curves of the four different approaches.

The classifier fusion largely improves the precision on all datasets, except for TA2 with the proposed method. An explanation of this behaviour is that the upper-body detections from the TA2 dataset are “cleaner” in terms of colour segmentation. Also, the false positives are easier to distinguish from true positives compared to the other datasets. Thus, the fusion with the detection score does not help in that case, but it does not decrease the precision neither. Also, one can note that the proposed multi-class segmentation algorithm is performing considerably better than the two-class segmentation proposed by Ramanan. This is obvious when the fusion step is not used, but also visible when exploiting it.

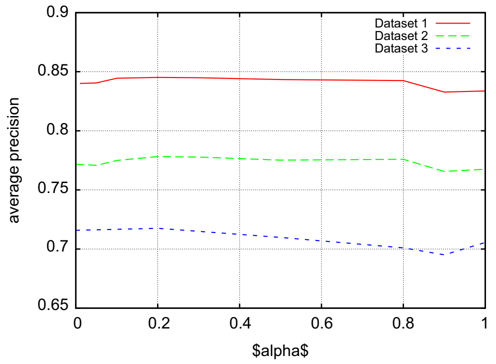

Fig. 9. Average precision for the three datasets with the Calvin detector and with varying parameter α, *i.e.* the proportion of the uniform distribution involved in the PIM class prior distribution (see Eq. (6)).

Also note that the results of Ramanan's method with classifier fusion are only marginally better than the base detector results, because the scores given by the colour segmentation classifier are much less reliable.

Uniform prior amount: We investigate the influence of the parameter α in Eq. (6) in the segmentation algorithm, *i.e.* the amount of the uniform PIM spatial prior with respect to the trained prior. Fig. 9 shows the average precision for the three datasets with the Calvin detector and with varying α. One can see that the use of a uniform prior is not crucial but slightly improves the average precision. The maximum lies around $\alpha = 0.2$, the value we used in our experiments.

Qualitative results: Finally, Fig. 10 illustrates some cases where the proposed algorithm, including segmentation and fusion, helped to improve the upper-body detection precision (top three rows), and reversely, where it had problems (bottom two rows). More results are given in the supplementary material. In general, we noticed that for samples where the approach improved classification, the shape classifier was often confused by the presence of high gradients at sensitive places, like within the face region due to strong shadows (first row of Fig. 10) or like the horizontal edges near the arm (second row) which overall form a configuration that can often be found in negative samples like buildings. The employed colour segmentation algorithm is able to filter out these confusing gradients and to recover and correctly identify only the relevant upper-body shape. Cases where the colour classifier degraded the score of true samples were often due to ambiguous clothing/background segmentation happening when the same colour is found in both regions (4th row). Finally, the score of negative samples happened to increase when the segmentation maps actually looked like human colour segmentation maps (5th row). Note that in this later case, even the observed colour fits the face and hair colour prior.

4. Conclusions

We presented a new method for upper-body detection that extends state-of-the-art detectors commonly based on shape features,

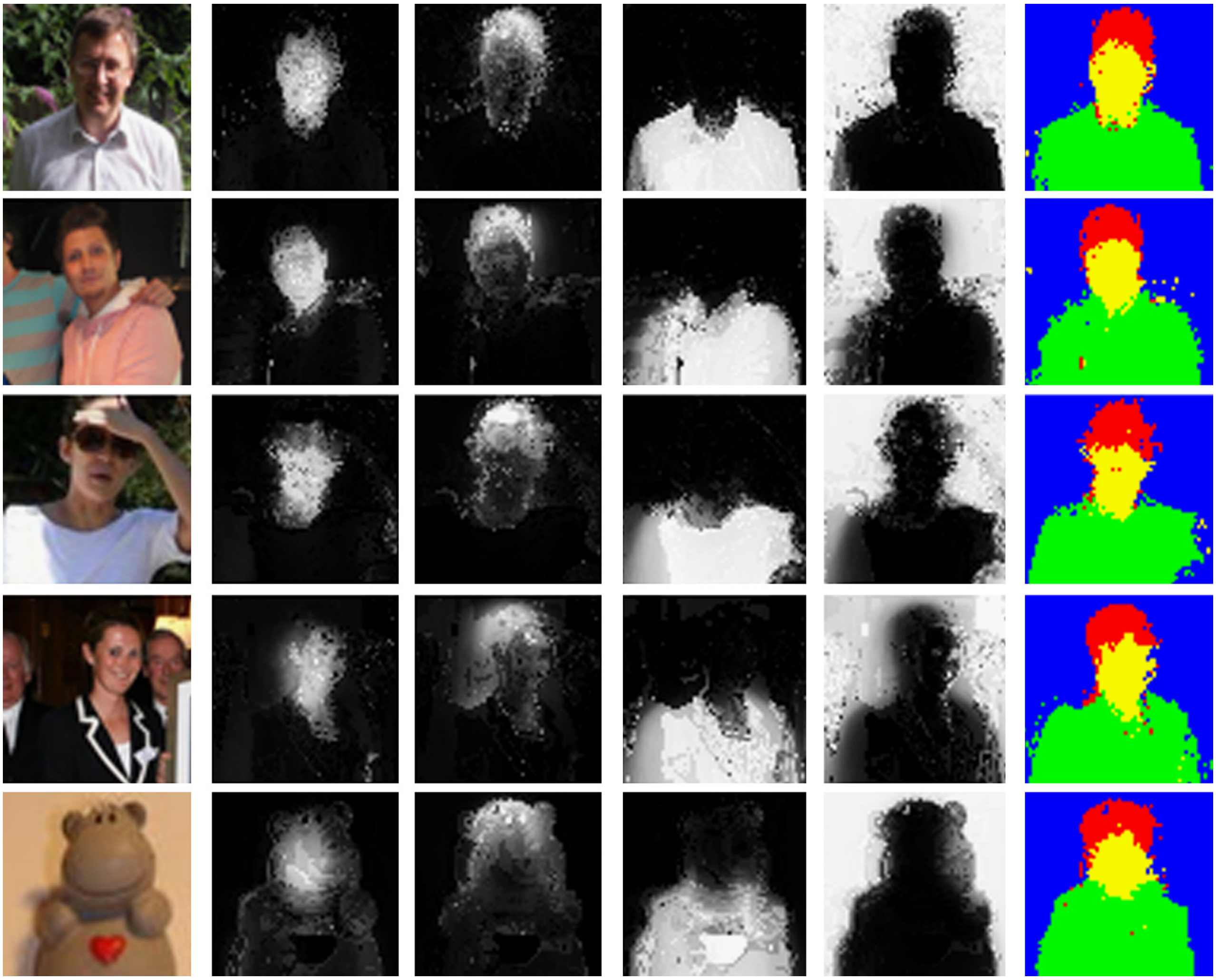

Fig. 10. Some examples where the proposed approach improved classification (top three rows) or degraded it (bottom two rows). First column: input image; 2nd to 5th columns: segmentation posteriors; last column: maximum *a posteriori* segmentation (which is not used by the algorithm). *Top three rows:* positive examples of upper-body detections where the rank from the final score has been significantly higher than the rank from the detector score. *Fourth row:* true detection with decreased rank due to similar clothing and background colour. *Bottom row:* negative sample where our method had problems, *i.e.* increased the rank. Note the human shape segmentation and the skin-like object colour which matches the face colour prior.

like HOG. The algorithm makes use of multi-class segmentation features extracted by a probabilistic colour segmentation approach. We showed experimentally that, by classifying the soft segmentations resulting from shape-based upper-body detections and fusing both detection and segmentation scores, the overall detection precision is improved. The proposed method outperforms state-of-the-art upper-body detectors based on shape features, suggesting that colour segmentation contains complementary, discriminatory information with respect to shape.

Future work could investigate the benefit of using this type of colour segmentation features for other visual object detection tasks. Also the effect of using of several shape priors to cope for different views and object poses could be explored. Further, the proposed detection and segmentation approach could be used for efficient object or person tracking, where the prior models could potentially be adapted over time. Finally, one might consider to directly include these segmentation features in the detector, and to train both shape and colour features in a common framework.

Conflict of interest

None declared.

Appendix A. Supplementary material

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2013.12.014.

References

[1] M. Eichner, M. Marin-Jimenez, A. Zisserman, V. Ferrari, Articulated Human Pose Estimation and Search in (Almost) Unconstrained Still Images, Technical Report 272, ETH Zurich, D-ITET, BIWI, 2010.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1627–1645.

[3] A. Mohan, C. Papageorgiou, T. Poggio, Example-based object detection in images by components, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 349–361.

[4] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, vol. 1.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 886–893.

[6] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Progressive search space reduction for human pose estimation, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–8.

[7] C. Hou, H. Ai, S. Lao, Multiview pedestrian detection based on vector boosting, in: ACCV.

Please cite this article as: S. Duffner, J.-M. Odobez, Leveraging colour segmentation for upper-body detection, Pattern Recognition (2014), http://dx.doi.org/10.1016/j.patcog.2013.12.014

[8] M. Li, Z. Zhang, K. Huang, T. Tan, Rapid and robust human detection based on omega-shape, in: ICIP.

[9] X. Ding, H. Xu, P. Cui, L. Sun, S. Yang, A cascade svm approach for head-shoulder detection using histograms of oriented gradients, in: IEEE International Symposium on Circuits and Systems, pp. 1791–1794.

[10] C. Zeng, H. Ma, Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting, in: Proceedings of the International Conference on Pattern Recognition, pp. 2069–2072.

[11] K. Mikolajczyk, C. Schmid, A. Zisserman, Human detection based on a probabilistic assembly of robust part detectors, in: Proceedings of the European Conference on Computer Vision, pp. 69–82.

[12] A.S. Micilotta, E.-J. Ong, R. Bowden, Real-time upper body detection and 3d pose estimation in monoscopic images, in: Proceedings of the European Conference on Computer Vision, pp. 139–150.

[13] R.Y.D. Xu, M. Kemp, Multiple curvature based approach to human upper body parts detection with connected ellipse model fine-tuning, in: Proceedings of the International Conference on Image Processing, pp. 2577–2580.

[14] P. Buehler, M. Everingham, D.P. Huttenlocher, A. Zisserman, Upper body detection and tracking in extended signing sequences, Int. J. Comput. Vis. 95 (2011) 180–197.

[15] P. Ott, M. Everingham, Implicit color segmentation features for pedestrian and object detection, in: Proceedings of the International Conference on Computer Vision, pp. 723–730.

[16] W.R. Schwartz, A. Kembhavi, D. Harwood, L.S. Davis, Human detection using partial least squares analysis, in: Proceedings of the International Conference on Computer Vision, pp. 24–31.

[17] D. Ramanan, Using segmentation to verify object hypotheses, in: CVPR.

[18] C. Scheffler, J. Odobez, Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps, in: Proceedings of the British Machine Vision Conference.

[19] N. Jojic, Y. Caspi, Capturing image structure with probabilistic index maps, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition.

[20] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, Int.J. Comput. Vis. 46 (2002) 81–96.

[21] I. Laptev, Improving object detection with boosted histograms, Image Vis. Comput. 27 (2009) 535–544.

[22] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.

Stefan Duffner received a Bachelors degree in Computer Science from the University of Applied Sciences Konstanz, Germany, in 2002 and a Masters degree in Applied Computer Science from the University of Freiburg, Germany, in 2004. He performed his dissertation research at Orange Labs in Rennes, France, on face image analysis with statistical machine learning methods, and in 2008, he obtained a Ph.D. degree in Computer Science from the University of Freiburg. He then worked for 4 years as a post-doctoral researcher at the Idiap Research Institute in Martigny, Switzerland, in the field of computer vision and mainly face tracking. As of today, he is an associate professor in the IMAGINE team of the LIRIS research lab at the National Institute of Applied Sciences (INSA) of Lyon, France.

Jean-Marc Odobez received an engineering degree from the Ecole Nationale Supérieure de Télécommunications de Bretagne (ENSTBr), Brest, France, in 1990, and the PhD. degree in signal processing from Rennes University, France, in 1994. He performed his dissertation research at IRISA/INRIA Rennes on dynamic scene analysis using statistical models. He then spent 1 year as a postdoctoral fellow at the GRASP Laboratory, University of Pennsylvania, working on visually guided robotic navigation problems. From 1996 until September 2001, he was an associate professor in computer science at the Université du Maine, Le Mans, France. He is now a senior researcher at both the IDIAP Research Institute and EPFL, Switzerland, where he directs the Human Activity analysis team. His main areas of research are computer vision and machine learning techniques applied to multimedia content analysis as well as tracking and human activity and behavior recognition. He is the author or coauthor of more than 100 papers in international journals and conferences in his field. He is or was the principle investigator of 10 European and Swiss projects. He holds two patents on video motion analysis. He is the cofounder of the Swiss Klewel SA company active in the intelligent capture, indexing, and Web casting of multimedia conference and seminar events. He is a member of the IEEE, and an associate editor of the Machine Vision and Application journal.