# Learning to Segment Humans by Stacking their Body Parts

E. Puertas, MA. Bautista, D. Sanchez, S. Escalera and O. Pujol

Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona,
Gran Via 585, 08007, Barcelona, Spain.
Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain
**{eloi,mabautista,dsanchez,sergio,oriol}@maia.ub.es**

**Abstract.** Human segmentation in still images is a complex task due to the wide range of body poses and drastic changes in environmental conditions. Usually, human body segmentation is treated in a two-stage fashion. First, a human body part detection step is performed, and then, human part detections are used as prior knowledge to be optimized by segmentation strategies. In this paper, we present a two-stage scheme based on Multi-Scale Stacked Sequential Learning (MSSL). We define an extended feature set by stacking a multi-scale decomposition of body part likelihood maps. These likelihood maps are obtained in a first stage by means of a ECOC ensemble of soft body part detectors. In a second stage, contextual relations of part predictions are learnt by a binary classifier, obtaining an accurate body confidence map. The obtained confidence map is fed to a graph cut optimization procedure to obtain the final segmentation. Results show improved segmentation when MSSL is included in the human segmentation pipeline.

**Keywords:** Human body segmentation, Stacked Sequential Learning

## 1 Introduction

Human segmentation in RGB images is a challenging task due to the high variability of the human body, which includes a wide range of human poses, lighting conditions, cluttering, clothes, appearance, background, point of view, number of human body limbs, etc. In this particular problem, the goal is to provide a complete segmentation of the person/people appearing in an image. In literature, human body segmentation is usually treated in a two-stage fashion. First, a human body part detection step is performed, obtaining a large set of candidate body parts. These parts are used as prior knowledge by segmentation/inference optimization algorithms in order to obtain the final human body segmentation.

In the first stage, that is the detection of body parts, weak classifiers are trained in order to obtain a soft prior of body parts (which are often noisy and unreliable). Most works in literature have used edge detectors, convolutions with filters, linear SVM classifiers, Adaboost or Cascading classifiers [27]. For example, [22] used a tubular edge template as a detector, and convolved it with

an image defining locally maximal responses above a threshold as detections. In [21], the authors used quadratic logistic regression on RGB features as the part detectors. Other works, have applied more robust part detectors such as SVM classifiers [5, 16] or AdaBoost [19] trained on HOG features [7]. More recently, Dantone et. al used Random Forest as classifiers to learn body parts [9]. Although recently robust classifiers have been used, part detectors still involve false-positive and false-negatives problems given the similarity nature among body parts and the presence of background artifacts. Therefore, a second stage is usually required in order to provide an accurate segmentation.

In the second stage, soft part detections are jointly optimized taking into account the nature of the human body. However, standard segmentation techniques (*i.e.*region-growing, thresholding, edge detection, etc.) are not applicable in this context due to the huge variability of environmental factors (i.e lightning, clothing, cluttering, etc.) and the changing nature of body textures. In this sense, the most known models for the optimization/inference of soft part priors are Poselets [4, 19] of Bourdev et. al. and Pictorial Structures [14, 2, 24] by Felzenszwalb et. al., both of which optimize the initial soft body part priors to obtain a more accurate estimation of the human pose, and provide with a multi-limb detection. In addition, there are some works in literature that tackle the problem of human body segmentation (segmenting the full body as one class) obtaining satisfying results. For instance, Vinet et al. [26] proposed to use Conditional Random Fields (CRF) based on body part detectors to obtain a complete person/background segmentation. Belief propagation, branch and bound or Graph Cut optimization are common approaches used to perform inference of the graphical models defined by human body [17, 23, 18]. Finally, methods like structured SVM or mixture of parts [29, 28] can be use in order to take profit of the contextual relations of body parts.

In this paper, we present a novel two-stage human body segmentation method based on the discriminative Multi-Scale Stacked Sequential Learning (MSSL) framework [15]. Until now stacked sequential learning has been used in several domains, mainly in text sequences and time series [6, 11] showing important computational and performance improvements when compared with other contextual inference methods such as CRF. Recently, the MSSL framework has been also successfully used on pixel wise classification problems [20]. To the best of our knowledge this is the first work that uses MSSL in order to find a context-aware feature set that encodes high order relations between body parts, which suffer non-rigid transformations, to obtain a robust human body segmentation. Fig. 1 shows the proposed human body segmentation approach. In the first stage of our method for human segmentation, a multi-class Error-Correcting Output Codes classifier (ECOC) is trained to detect body parts and to produce a soft likelihood map for each body part. In the second stage, a multi-scale decomposition of these maps and a neighborhood sampling is performed, resulting in a new set of features. The extended set of features encodes spatial, contextual and relational information among body parts. This extended set is then fed to the second classifier of MSSL, in this case a Random Forest binary classifier,

which maps a multi-limb classification to a binary human classification problem. Finally, in order to obtain the resulting binary human segmentation, a post-processing step is performed by means of Graph Cuts optimization, which is applied to the output of the binary classifier.

The rest of the paper is organized as follows: Section 2 introduces the proposed method. Section 3 presents the experimental results. Finally, Section 4 concludes the paper.

## 2 Method

The proposed method for human body segmentation is based on the Multi-Scale Stacked Sequential Learning (MSSL)[15] pipeline. Generalized Stacked Sequential Learning was proposed as a method for solving the main problems of sequential learning, namely: (a) how to capture and exploit sequential correlations; (b) how to represent and incorporate complex loss functions in contextual learning; (c) how to identify long-distance interactions; and (d) how to make sequential learning computationally efficient. Fig. 1 (a) shows the abstract blocks of the process[1]. Consider a training set consisting of data pairs $\{(x_i, y_i)\}$, where $x_i \in \mathcal{R}^n$ is a feature vector and $y_i \in \mathcal{Y}, \quad \mathcal{Y} = \{1, \ldots, K\}$ is the class label. The first block of MSSL consists of a classifier $H_1(x)$ trained with the input data set. Its output results are a set of predicted labels or confidence values $Y'$. The next block in the pipeline, defines the policy for taking into account the context and long range interactions. It is composed of two steps: first, a multi-resolution decomposition models the relationship among neighboring locations, and second, a neighborhood sampling proportional to the resolution scale defines the support lattice. This last step allows to model the interaction range. This block is represented by the function $z = J(x, \rho, \theta) : \mathcal{R} \rightarrow \mathcal{R}^w$, parameterized by the interaction range $\theta$ in a neighborhood $\rho$. The last step of the algorithm creates an extended data set by adding to the original data the new set of features resulting from the sampling of the multi resolution confidence maps which is the input of a second classifier $H_2(x)$.

### 2.1 Stage One: Body Parts Soft Detection

In this work, the first stage detector $H_1(x)$ in the MSSL pipeline is based on the soft body parts detectors defined in [8]. The work of Bautista *et al.* [8] is based on an ECOC ensemble of cascades of Adaboost classifiers. Each of the cascades focuses on a subset of body parts described using Haar-like features where regions have been previously rotated towards main orientation to make the recognition rotation invariant. Although any other part detector technique could be used in the first stage of our process, we also choose the same methodology. ECOC has shown to be a powerful and general framework that allows the inclusion of any

[1] The original formulation of MSSL also includes the input vector $X$ as an additional feature in the extended set $X'$.

base classifier, involving error-correction capabilities and allowing to reduce the bias and variance errors of the ensemble [10, 12]. As a case study, although any classifier can be included in the ECOC framework, here we considerer as base learner also the same ensemble of cascades given its fast computization.

Because of its properties, cascades of classifiers are usually trained to split one visual object from the rest of possible objects of an image. This means that the cascade of classifiers learns to detect a certain object (body part in our case), ignoring all other objects (all other body parts). However, some body parts have similar appearance, *i.e.*legs and arms, and thus, it makes sense to group them in the same visual category. Because of this, we learn a set of cascades of classifiers where a subset of limbs are included in the positive set of one cascade, and the remaining limbs are included as negative instances together with background images in the negative set of the cascade. In this sense, classifier $H_1$ is learned by grouping different cascades of classifiers in a tree-structure way and combining them in an Error-Correcting Output Codes (ECOC) framework [13]. Then, $H_1$ outputs correspond to a multi-limb classification prediction.

An example of the body part tree-structure defined taking into account the nature of human body parts is shown in Fig. 2(a). Notice that classes with similar visual appearance (*e.g.*upper-arm and lower-arm) are grouped in the same meta-class in most dichotomies. In addition, dichotomies that deal with difficult problems (*e.g.*$d^5$) are focused only in the difficult classes, without taking into account all other body parts. In this case, class $c^7$ denotes the background.

In the ECOC framework, given a set of $K$ classes (body parts) to be learnt, $m$ different bi-partitions (groups of classes or dichotomies) are formed, and $n$ binary problems over the partitions are trained [3]. As a result, a codeword of length $n$ is obtained for each class, where each position (bit) of the code corresponds to a response of a given classifier $d$ (coded by $+1$ or $-1$ according to their class set membership, or 0 if a particular class is not considered for a given classifier). Arranging the codewords as rows of a matrix, we define a *coding matrix* $M$, where $M \in \{-1, 0, +1\}^{K \times n}$. During the *decoding* (or testing) process, applying the $n$ binary classifiers, a code $c$ is obtained for each data sample $x$ in the test set. This code is compared to the base codewords ($y^i, i \in \{1, .., K\}$[2]) of each class defined in the matrix $M$, and the data sample is assigned to the class with the *closest* codeword [13].

We use the problem dependent coding matrix defined in [8] in order to allow the inclusion of cascade of classifiers and learn the body parts. In particular, each dichotomy is obtained from the body part tree-structure. Fig. 2(b) shows the coding matrix codification of the tree-structure in Fig. 2(a).

In the ECOC *decoding* step an image is processed using a sliding windowing approach. Each image patch $x$, is described and tested. In our case, each patch is first rotated by main gradient orientation and tested using the ECOC ensemble with Haar-like features and cascade of classifier. In this sense, each classifier $d$

[2] Observe that we are overloading the notation of $y$ so that $y^i$ corresponds to the codeword of the matrix associated with class $i$, *i.e.*it is the i-th row of the matrix, $M(i, :)$.

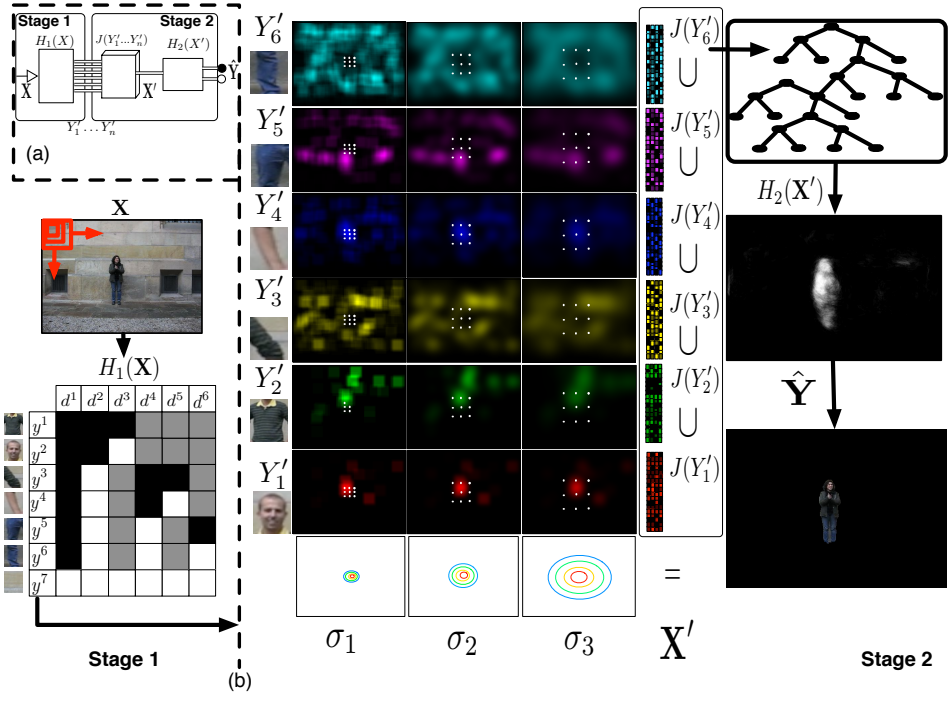



**Fig. 1.** Method overview. (a) Abstract pipeline of the proposed MSSL method where the outputs $Y'_i$ of the first multi-class classifier $H_1(x)$ are fed to the multi-scale decomposition and sampling function $J(x)$ and then used to train the second stacked classifier $H_2(x)$ which provides a binary output $\hat{\mathbf{Y}}$. (b) Detailed pipeline for the MSSL approach used in the human segmentation context where $H_1(x)$ is a multi-class classifier that takes a vector $\mathbf{X}$ of images from a dataset. As a result, a set of likelihood maps $Y'_1 \dots Y'_n$ for each part is produced. Then a multi-scale decomposition with a neighborhood sampling function $J(x)$ is applied. The output $\mathbf{X}'$ produced is taken as the input of the second classifier $H_2(x)$, which produces the final likelihood map $\hat{\mathbf{Y}}$, showing for each point the confidence of belonging to human body class.

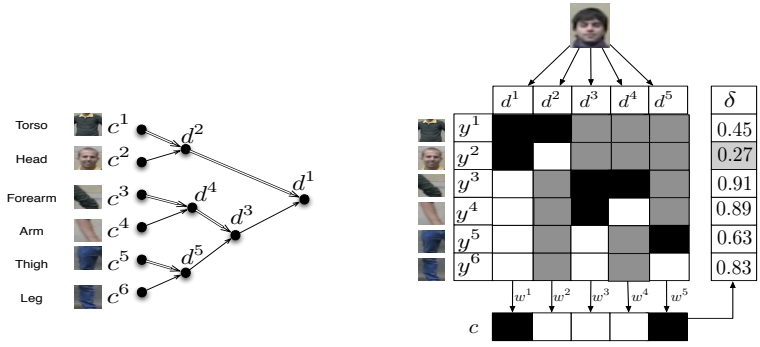
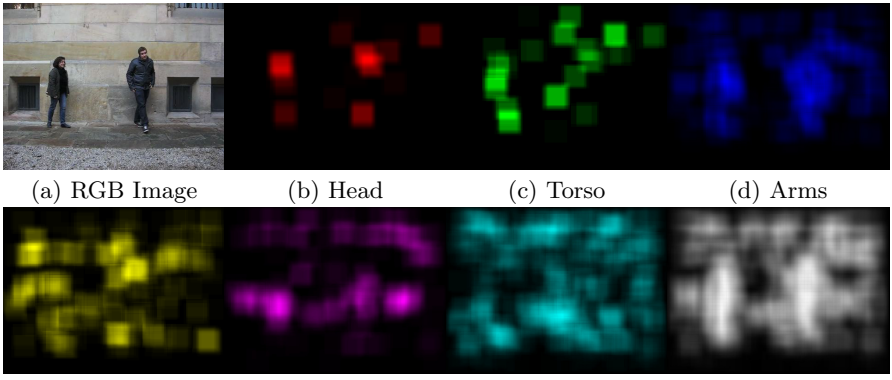



**Fig. 2.** (a) Tree-structure classifier of body parts, where nodes represent the defined dichotomies. Notice that the single or double lines indicate the meta-class defined. (b) ECOC decoding step, in which a head sample is classified. The coding matrix codifies the tree-structure of (a), where black and white positions are codified as $+1$ and $-1$, respectively. $c$, $d$, $y$, $w$, $X$, and $\delta$ correspond to a class category, a dichotomy, a class codeword, a dichotomy weight, a test codeword, and a decoding function, respectively.

outputs a prediction whether $x$ belongs to one of the two previously learnt meta-classes. Once the set of predictions $c \in \{+1, -1\}^{1\times n}$ is obtained, it is compared to the set of codewords of the classes $y^i$ from $M$, using a decoding function $\delta(c, y^i)$ and the final prediction is the class with the codeword with minimum decoding, *i.e.*$\arg\min_i \delta(c, y^i)$. As a decoding function we use the Loss-Weighted approach with linear loss function defined in [13]. Then, a body-like probability map is built. This map contains, at each position the proportion of body part detections for each pixel over the total number of detections for the whole image. In other words, pixels belonging to the human body will show a higher body-like probability than the pixels belonging to the background. Additionally, we also construct a set of limb-like probability maps. Each map contains at each position $(i, j)$ the probability of pixel at the entry $(i, j)$ of belonging to the body part class. This probability is computed as the proportion of detections at point $(i, j)$ over all detection for that class. Examples of probability maps obtained from ECOC outputs are shown in Fig. 3, which represents the $H_1(x)$ outputs $Y'_1 \dots Y'_n$ defined in Fig. 1 (a).

### 2.2 Stage Two: Fusing Limb Likelihood Maps Using MSSL

The goal of this stage is to fuse all partial body parts into a full human body likelihood map (see Fig. 1 (b) second stage). The input data for the neighborhood modeling function $J(x)$ are the body parts likelihood maps obtained in the first stage ($Y'_1 \dots Y'_n$). In the first step of the modeling a set of different gaussian filters is applied on each map. All these multi-resolution decompositions give information about the influence of each body part at different scales along the



(a) RGB Image (b) Head (c) Torso (d) Arms

(e) Forearms (f) Thighs (g) Legs (d) Full Body

**Fig. 3.** Limb-like probability maps for the set of 6 limbs and body-like probability map. Image (a) shows the original RGB image. Images from (b) to (g) illustrate the limb-like probability maps and (h) shows the union of these maps.

space. Then, a 8-neighbor sampling is performed for each pixel with sampling distance proportional to its decomposition scale. This allows to take into account the different limbs influence and their context. The extended set $X'$ is formed by stacking all the resulting samplings at each scale for each limb likelihood map (see the extended feature set $X'$ in Fig. 1(b)). As a result, $X'$ will have dimensionality equals to the number of samplings multiplied by the number of scales and the number of body parts. In our experiments we use eight neighbor sampling, three scales and six body parts. Notice that contrary to the MSSL traditional framework, we do not fed the second classifier $H_2$ with both the original $X$ and extended $X'$ features, and only the extended set $X'$ is provided. In this sense, the goal of $H_2$ is to learn spatial relations among body parts based on the confidences produced by first classifier. As a result, second classifier provides a likelihood of the membership of an image pixel to the class 'person'. Thus, the multiple spatial relations of body parts (obtained as a multi-class classifier in $H_1$), are labelled as a two-class problem (*person* vs *not person*) and trained by $H_2$. Consequently, the label set associated to the extended training data $X'$ corresponds to the union of the ground truths of all human body parts. Although, within our method any binary classifier can be considerer for $H_2$, we use a Random Forest classifier to train 50 random trees that focus on different configurations of the data features. This strategy has shown robust results for human body segmentation in multi-modal data [25]. Fig. 4 shows a comparative between the union of the likelihood maps obtained by the first classifier and the final likelihoods obtained after the second stage. We can see that a naive fusion of the limb likelihoods produce noisy outputs in many body parts. The last column shows how second stage clearly detects the human body using the same data. For instance, Fig. 4 (f) shows how it works well also when two bodies are close one to other, splitting them accurately, preserving the poses. Notice

that in Fig. 4 (f) a non zero probability zone exists between both silhouettes, denoting the existence of a handshaking. Finally in Fig. 4 (c) we can see how the foreground person is highlighted in the likelihood map, while in previous stage (Fig. 4 (b)) it was completely missed. This shows that the second stage is able to restore body objects at different scales. Finally, the output likelihood maps obtained after this stage are used as input of a post-process based on graph-cut to obtain final segmentation

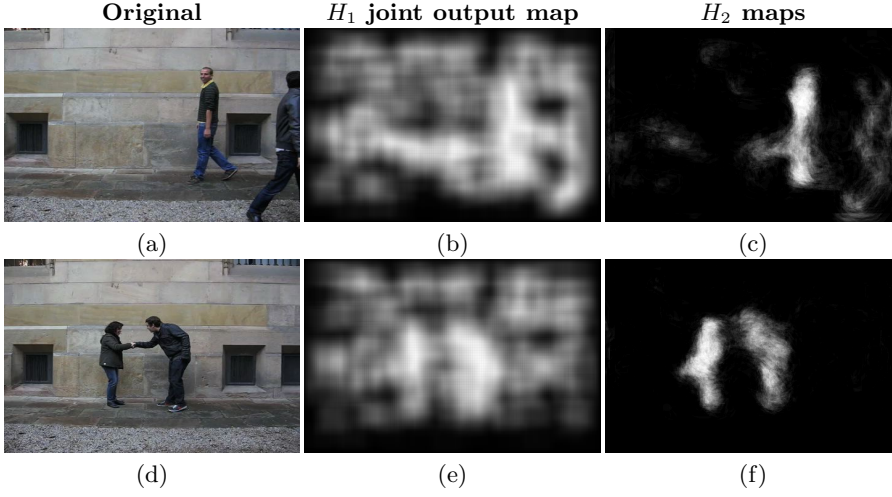



**Fig. 4.** Comparative between $H_1$ and $H_2$ output. First column are the original images. Second column are $H_2$ output likelihood maps. Last column are the union of all likelihood map of body parts

## 3 Experimental Results

Before present the experimental results, we first discuss the data, experimental settings, methods and validation protocol.

### 3.1 Dataset

We used *HuPBA* $8k+$ *dataset* described in [1]. This dataset contains more than 8000 labeled images at pixel precision, including more than 120000 manually labeled samples of 14 different limbs. The images are obtained from 9 videos (RGB sequences) and a total of 14 different actors appear in those 9 sequences. In concrete, each sequence has a main actor (9 in total) which during the sequence interacts with secondary actors portraying a wide range of poses. For our experiments, we reduced the number of limbs from the 14 available in the

dataset to 6, grouping those that are similar by symmetry (right-left) as arms, forearms, thighs and legs. Thus, the set of limbs of our problem is composed by: *head*, *torso*, *forearms*, *arms*, *thighs* and *legs*. Although labeled within the dataset, we did not include hands and feet in our segmentation scheme. In Fig. 5 some samples of the *HuPBA* $8k+$ dataset are shown.

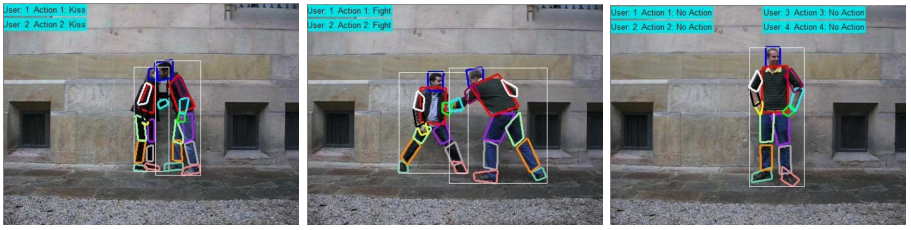



**Fig. 5.** Different samples of the HuPBA $8k+$ dataset.

### 3.2 Methods

We compare the following methods for Human Segmentation: **Soft Body Parts (SBP) detectors + MSSL + Graphcut**. The proposed method, where the body like confidence map obtained by each body part soft detector is learned by means of MSSL and the output is then fed to a GraphCut optimization to obtain the final segmentation. **SBP detectors + MSSL + GMM-Graphcut**. Variation of the proposed method, where the final GraphCut optimization also learns a GMM color model to obtain the final segmentation as in the GrabCut model [23]. **SBP detectors + GraphCut**. In this method the body like confidence map obtained by aggregating all body parts soft detectors outputs is fed to a GraphCut optimization to obtain the final segmentation. **SBP detectors + GMM-GraphCut**. We also use the GMM color modeling variant in the comparison.

### 3.3 Settings and validation protocol

In a preprocessing step, we resized all limb samples to a $32 \times 32$ pixels region. Regions are first rotated by main gradient orientation. In the first stage, we used the standard Cascade of Classifiers based on AdaBoost and Haar-like features [27] as our body part multi-class classifier $H_1$. As model parameters, we forced a 0.99 false positive rate and maximum of 0.4 false alarm rate during 8 stages. To detect limbs with trained cascades of classifiers, we applied a sliding window approach with an initial patch size of $32 \times 32$ pixels up to $60 \times 60$ pixels. As result of this stage, we obtained 6 likelihood maps for each image. In the second stage, we performed 3-scale gaussian decomposition with $\sigma \in [8, 16, 32]$ for each body part. Then, we generated a extended set selecting for each pixel its 8-neighbors with $\sigma$ displacement. From this extended set, a sampling of 1500 selected points formed

the input examples for the second classifier. As second classifier, we used a Random Forest with 50 decision trees. Finally, in a post-processing stage, binary Graph Cuts with a GMM color modeling (we experimentally set 3 components) were applied to obtain the binary segmentation where the initialization seeds of foreground and background were tuned via cross-validation. For the binary Graph Cuts without a GMM color modeling we directly fed the body likelihood map to the optimization method. In order to assess our results, we used 9-fold cross-validation, where each fold correspond to images of a main actor sequence. As results measurement we used the Jaccard Index of overlapping ($J = \frac{A \bigcap B}{A \bigcup B}$) where $A$ is the ground-truth and $B$ is the corresponding prediction.

### 3.4 Quantitative Results

In Table 1 we show overlapping results for the *HuPBA* $8K+$ dataset. Specifically, we show the mean overlapping value obtained by the compared methods on 9 folds of the *HuPBA* $8k+$ dataset. We can see how our MSSL proposal consistently obtains a higher overlapping value on every fold.

| | **GMM-GC** | | **GC** | |
|---|---|---|---|---|
| | **MSSL** | **Soft Detect.** | **MSSL** | **Soft Detect.** |
| **Fold** | **Overlap** | **Overlap** | **Overlap** | **Overlap** |
| 1 | **62.35** | 60.35 | **63.16** | 60.53 |
| 2 | **67.77** | 63.72 | **67.28** | 63.75 |
| 3 | **62.22** | 60.72 | **61.76** | 60.67 |
| 4 | **58.53** | 55.69 | **58.28** | 55.42 |
| 5 | **55.79** | 51.60 | **55.21** | 51.53 |
| 6 | **62.58** | 56.56 | **62.33** | 55.83 |
| 7 | **63.08** | 60.67 | **62.79** | 60.62 |
| 8 | **67.37** | 64.84 | **67.41** | 65.41 |
| 9 | **64.95** | 59.83 | **64.21** | 59.90 |
| Mean | **62,73** | 59,33 | **62,49** | 59,29 |

**Table 1.** Overlapping results over the 9 folds of the *HupBA8K+* dataset for the proposed MSSL method and the Soft detectors post-processing their outputs with the Graph-Cuts method and GMM Graph-Cuts method.

Notice that MSSL proposal outperforms in the SBP+GC method in all folds (by at least a 3% difference), which is the state-of-the-art method for human segmentation in the HuPBA $8k+$ dataset [8].

### 3.5 Qualitative Results

In Fig. 6 some qualitative results of the compared methodologies for human segmentation are shown. It can be observed how in general SBP+MSSL+GMM-GC obtains a better segmentation of the human body than the SBP + GMM-GC method. This improvement is due to the contextual body part information

encoded in the extended feature set. In particular, this performance difference is clearly visible in Fig. 6(f) where the human pose is completely extracted from the background. We also observe how the proposed method is able to detect a significative number of body parts at different scales. This is clearly appreciated in Fig. 6(c), where persons at different scales are segmented, while in Fig. 6(b) the SBP+GMM-GC fails to segment the rightmost person. Furthermore, Fig. 6(i) shows how the proposed method is able to recover the whole body pose by stacking all body parts, while in Fig. 6(h) the SBP+GMM-GC method just detected the head of the left most user. In this pair of images also we can see how our method is able to discriminate the different people appearing in an image, segmenting as background the interspace between them. Although, it may cause some loss, specially in the thinner body parts, like happens with the extended arm. Due to space restrictions, a table with more examples of segmentation results can be found in the supplementary material. Regards the dataset used, it is important to remark the large amount of segmented bodies (more than 10.000) and their high variability in terms of pose (performing different activities and interactions with different people), size and clothes. The scale variations are learnt by $H_2$ through spatial relationships of body parts. In addition, although background is maintained across the data, $H_2$ is trained over the soft predictions from $H_1$ (see the large number of false positive predictions shown in Fig. 3), and our method considerably improves those person confidence maps, as shown in Fig. 4.

## 4 Conclusions

We presented a two-stage scheme based on the MSSL framework for the segmentation of the human body in still images. We defined an extended feature set by stacking a multi-scale decomposition of body part likelihood maps, which are learned by means of a multi-class classifier based on soft body part detectors. The extended set of features encodes spatial and contextual information of human limbs which combined enabled us to define features with high order information. We tested our proposal on a large dataset obtaining significant segmentation improvement over state-of-the-art methodologies. As future work we plan to extend the MSSL framework to the multi-limb case, in which two multi-class classifiers will be concatenated to obtain a multi-limb segmentation of the human body that takes into account contextual information of human parts.

## References

1. http://gesture.chalearn.org/. Tech. rep.
2. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1014–1021. IEEE (2009)
3. Bautista, M.A., Escalera, S., Baró, X., Radeva, P., Vitriá, J., Pujol, O.: Minimal design of error-correcting output codes. Pattern Recogn. Lett. 33(6), 693–702 (Apr 2012)

4. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Computer Vision–ECCV 2010, pp. 168–181. Springer (2010)
5. Chakraborty, B., Bagdanov, A.D., Gonzalez, J., Roca, X.: Human action recognition using an ensemble of body-part detectors. Expert Systems (2011)
6. Cohen, W.W., de Carvalho, V.R.: Stacked sequential learning. Proc. of IJCAI 2005 pp. 671–676 (2005)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886 –893 vol. 1 (2005)
8. Daniel Sanchez, Juan Carlos Ortega, M.A.B., Escalera, S.: Human body segmentation with multi-limb error-correcting output codes detection and graph cuts optimization. In: Proceedings of InPRIA. pp. 50–58 (2013)
9. Dantone, M., Gall, J., Leistner, C., van Gool, L.: Human pose estimation using body parts dependent joint regressors. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 3041–3048 (June 2013)
10. Dietterich, T., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. In: Journal of Artificial Intelligence Research. vol. 2, pp. 263–286 (1995)
11. Dietterich, T.G.: Machine learning for sequential data: A review. Proc. on Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science pp. 15–30 (2002)
12. Escalera, S., Tax, D., Pujol, O., Radeva, P., Duin, R.: Subclass problem-dependent design of error-correcting output codes. PAMI 30(6), 1–14 (2008)
13. Escalera, S., Pujol, O., Radeva, P.: On the decoding process in ternary error-correcting output codes. PAMI 32, 120–134 (2010)
14. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient matching of pictorial structures. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. vol. 2, pp. 66–73. IEEE (2000)
15. Gatta, C., Puertas, E., Pujol, O.: Multi-scale stacked sequential learning. Pattern Recognition 44(10-11), 2414–2426 (2011)
16. Gkioxari, G., Arbelaez, P., Bourdev, L.D., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. In: CVPR. pp. 3342–3349. IEEE (2013)
17. Hernández-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., Escalera, S.: Graph cuts optimization for multi-limb human segmentation in depth maps. In: CVPR. pp. 726–732 (2012)
18. Hernández-Vela, A., Reyes, M., Ponce, V., Escalera, S.: Grabcut-based human segmentation in video sequences. Sensors 12(11), 15376–15393 (2012)
19. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 588–595. IEEE (2013)
20. Puertas, E., Escalera, S., Pujol, O.: Generalized multi-scale stacked sequential learning for multi-class classification. Pattern Analysis and Applications pp. 1–15 (2013)
21. Ramanan, D., Forsyth, D., Zisserman, A.: Strike a pose: tracking people by finding stylized poses. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 271–278 vol. 1 (June 2005)
22. Ramanan, D., Forsyth, D., Zisserman, A.: Tracking people by learning their appearance. PAMI 29(1), 65 –81 (jan 2007)
23. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23(3), 309–314 (Aug 2004)

24. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 422–429. IEEE (2010)
25. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: In In CVPR, 2011. 3 (2011)
26. Vineet, V., Warrell, J., Ladicky, L., Torr, P.: Human instance segmentation from video using detector-based conditional random fields. In: BMVC (2011)
27. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. vol. 1 (2001)
28. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1385–1392. IEEE (2011)
29. Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1169–1176. ACM (2009)

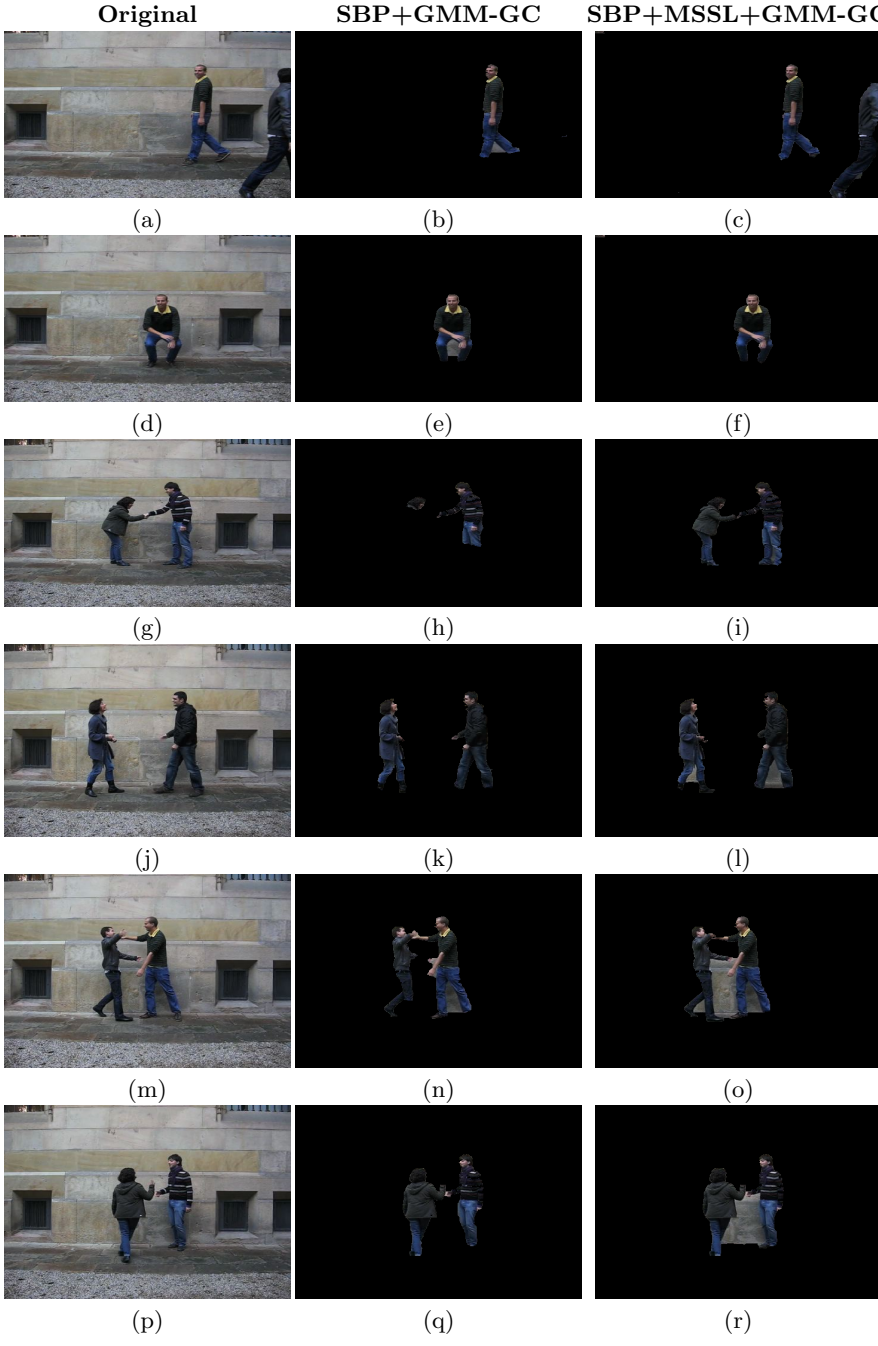



**Fig. 6.** Samples of the segmentation results obtained by the compared approaches.