Sensors **2012**, *12*, 15376-15393; doi:10.3390/s121115376

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

GrabCut-Based Human Segmentation in Video Sequences

Antonio Hernández-Vela [1,2,⋆], **Miguel Reyes** [1,2], **Víctor Ponce** [1,2] **and Sergio Escalera** [1,2]

[1] Departamento MAIA, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain;
E-Mails: mreyes@cvc.uab.cat (M.R.); vponce@cvc.uab.cat (V.P.); sergio@maia.ub.es (S.E.)

[2] Centre de Visió per Computador, Campus UAB, Edifici O, 08193 Bellaterra, Barcelona, Spain

***** Author to whom correspondence should be addressed; E-Mail: ahernandez@cvc.uab.cat;
Tel.: +34-93-402-1897; Fax: +34-93-402-1601.

Received: 4 September 2012; in revised form: 1 November 2012 / Accepted: 6 November 2012 / Published: 9 November 2012

Abstract: In this paper, we present a fully-automatic Spatio-Temporal GrabCut human segmentation methodology that combines tracking and segmentation. GrabCut initialization is performed by a HOG-based subject detection, face detection, and skin color model. Spatial information is included by Mean Shift clustering whereas temporal coherence is considered by the historical of Gaussian Mixture Models. Moreover, full face and pose recovery is obtained by combining human segmentation with Active Appearance Models and Conditional Random Fields. Results over public datasets and in a new Human Limb dataset show a robust segmentation and recovery of both face and pose using the presented methodology.

Keywords: segmentation; human pose recovery; GrabCut; GraphCut; Active Appearance Models; Conditional Random Field

1. Introduction

Human segmentation in uncontrolled environments is a hard task because of the constant changes produced in natural scenes: illumination changes, moving objects, changes in the point of view, occlusions, just to mention a few. Because of the nature of the problem, a common way to proceed is to discard most part of the image so that the analysis can be performed on a reduced set of small candidate regions. In [1], the authors propose a full-body detector based on a cascade of classifiers [2] using HOG features. This methodology is currently being used in several works related to the pedestrian detection problem [3]. GrabCut [4] has also shown high robustness in Computer Vision segmentation problems, defining the pixels of the image as nodes of a graph and extracting foreground pixels via iterated Graph Cut optimization. This methodology has been applied to the problem of human body segmentation with high success [5,6]. In the case of working with sequences of images, this optimization problem can also be considered to have temporal coherence. In the work of [7], the authors extended the Gaussian Mixture Model (GMM) of GrabCut algorithm so that the color space is complemented with the derivative in time of pixel intensities in order to include temporal information in the segmentation optimization process. However, the main problem of that method is that moving pixels corresponds to the boundaries between foreground and background regions, and thus, there is no clear discrimination.

Once a region of interest is determined, pose is often recovered by the determination of the body limbs together with their spatial coherence (also with temporal coherence in case of image sequences). Most of these approaches are probabilistic, and features are usually based on edges or "appearance". In [8], the author propose a probabilistic approach for limb detection based on edge learning complemented with color information. The image of probabilities is then formulated in a Conditional Random Field (CRF) scheme and optimized using belief propagation. This work has obtained robust results and has been extended by other authors including local GrabCut segmentation and temporal refinement of the CRF model [5,6].

In this paper, we propose a full-automatic Spatio-Temporal GrabCut human segmentation methodology, which benefits from the combination of tracking and segmentation. First, subjects are detected by means of a HOG-based cascade of classifiers. Face detection and skin color model are used to define a set of seeds used to initialize GrabCut algorithm. Spatial information is taken into account by means of Mean Shift clustering, whereas temporal information is considered taking into account the pixel probability membership to an historical of Gaussian Mixture Models. Moreover, the methodology is combined with Shape and Active Appearance Models (AAM) to define three different meshes of the face, one near frontal view, and the other ones near lateral views. Temporal coherence and fitting cost are considered in conjunction with GrabCut segmentation to allow a smooth and robust face fitting in video sequences. Finally, the limb detection and a CRF model are applied on the obtained segmentation, showing high robustness capturing body limbs due to the accurate human segmentation. The main limitation of our approach is that it depends on a correct detection of the person and his/her face, in order to get the desired result. In order to test the proposed methodology, we use public datasets and present a new Human Limb dataset useful for human segmentation, limb detection, and pose recovery purposes.

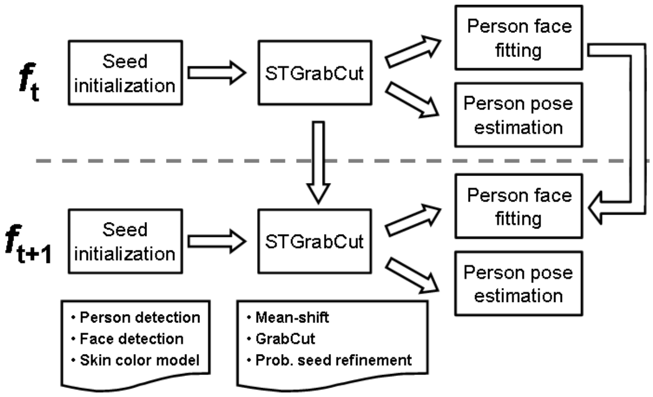
The rest of the paper is organized as follows: Section 2 describes the proposed methodology, presenting the spatio-temporal GrabCut segmentation, the AAM for face fitting, and the pose recovery methodology. Experimental results on public and novel datasets are performed in Section 3. Finally, Section 4 concludes the paper.

2. Full-Body Pose Recovery

In this section, we present the Spatio-Temporal GrabCut methodology to deal with the problem of automatic human segmentation in video sequences. Then, we describe the Active Appearance Models used to recover the face, and the body pose recovery methodology based on the approach of [8].

All methods presented in this section are combined to improve final segmentation and pose recovery. Figure 1 illustrates the different modules of the project.

Figure 1. Overall block diagram of the methodology.



2.1. GrabCut Segmentation

In [4], the authors proposed an approach to find a binary segmentation(background and foreground) of an image by formulating an energy minimization scheme as the one presented in [9–11], extended using color instead of just gray-scale information. Given a color image I, let us consider the array $z = (z_1, ..., z_n, ..., z_N)$ of N pixels where $z_i = (R_i, G_i, B_i)$, $i \in [1, ..., N]$ in RGB space. The segmentation is defined as array $\boldsymbol{\alpha} = (\alpha_1, ...\alpha_N)$, $\alpha_i \in \{0, 1\}$, assigning a label to each pixel of the image indicating if it belongs to background or foreground. A trimap T is defined by the user—in a semi-automatic way—consisting of three regions: T_B, T_F and T_U, each one containing initial background, foreground, and uncertain pixels, respectively. Pixels belonging to T_B and T_F are clamped as background and foreground respectively—which means GrabCut will not be able to modify these labels, whereas those belonging to T_U are actually the ones the algorithm will be able to label. Color information is introduced by GMMs. A full covariance GMM of K components is defined for background pixels ($\alpha_i = 0$), and another one for foreground pixels ($\alpha_j = 1$), parametrized as follows

$$\boldsymbol{\theta} = \{\pi(\alpha, k), \mu(\alpha, k), \Sigma(\alpha, k), \alpha \in \{0, 1\}, k = 1..K\}, \tag{1}$$

being π the weights, μ the means and Σ the covariance matrices of the model. We also consider the array $\mathbf{k} = \{k_1, ..., k_i, ...k_N\}$, $k_i \in \{1, ...K\}$, $i \in [1, ..., N]$ indicating the component of the background or foreground GMM (according to α_i) the pixel z_i belongs to. The energy function for segmentation is then

$$\mathbf{E}(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}) = \mathbf{U}(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}) + \mathbf{V}(\boldsymbol{\alpha}, \mathbf{z}), \tag{2}$$

where $\mathbf{U}$ is the likelihood potential, based on the probability distributions $p(\cdot)$ of the GMM:

$$\mathbf{U}(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}) = \sum_i -\log p(z_i|\alpha_i, k_i, \boldsymbol{\theta}) - \log \pi(\alpha_i, k_i) \tag{3}$$

and V is a regularizing prior assuming that segmented regions should be coherent in terms of color, taking into account a neighborhood C around each pixel

$$\mathbf{V}(\boldsymbol{\alpha}, \mathbf{z}) = \gamma \sum_{\{m,n\} \in C} [\alpha_n \neq \alpha_m] \exp\left(-\beta \|z_m - z_n\|^2\right) \tag{4}$$

With this energy minimization scheme and given the initial trimap T, the final segmentation is performed using a minimum cut algorithm [9,10,12]. The classical semi-automatic GrabCut algorithm is summarized in Algorithm 1.

Algorithm 1 Original GrabCut algorithm.

```
1: Trimap $T$ initialization with manual annotation.
2: Initialize $\alpha_i = 0$ for $i \in T_B$ and $\alpha_i = 1$ for $i \in T_U \cup T_F$.
3: Initialize Background and Foreground GMMs from sets $\alpha_i = 0$ and $\alpha_i = 1$ respectively, with
   $k$-means.
4: Assign GMM components to pixels.
5: Learn GMM parameters from data z.
6: Estimate segmentation: Graph-cuts.
7: Repeat from step 4, until convergence.
```

2.2. Automatic Initialization

Our proposal is based on the previous GrabCut framework, focusing on human body segmentation, being fully automatic, and extending it by taking into account temporal coherence. We refer to each frame of the video as f_t, $t \in \{1, ..., M\}$ being M the length of the sequence. Given a frame f_t, we first apply a person detector based on a cascade of classifiers using HOG features [1]. Then, we initialize the trimap T from the bounding box B retuned by the detector: $T_U = \{z_i \in B\}$, $T_B = \{z_i \notin B\}$. Furthermore, in order to increase the accuracy of the segmentation algorithm, we include Foreground seeds exploiting spatial and appearance prior information. On one hand, we define a small central rectangular region R inside B, proportional to B in such a way that we are sure it corresponds to the person. Thus, pixels inside R are set to foreground. On the other, we apply a face detector based on a cascade of classifiers using Haar-like features [2] over B, and learn a skin color model h_{skin} consisting of a histogram over the *Hue* channel of the *HSV* image representation. All pixels inside B fitting in h_{skin} are also set to foreground. Therefore, we initialize $T_F = \{z_i \in R\} \cup \{z_i \in \delta(z_i, h_{skin})\}$, where δ returns the set of pixels belonging to the color model defined by h_{skin}. An example of seed initialization is shown in Figure 2(b).

2.3. Spatial Extension

Once we have initialized the trimap, we can apply the iterative minimization algorithm shown in steps 4 to 7 of original GrabCut (Algorithm 1). However, instead of applying k-means for the initialization of the GMMs we propose to use Mean-Shift clustering, which also takes into account spatial coherence. Given an initial estimation of the distribution modes $m_h(\mathbf{x}^0)$ and a kernel function g, Mean-shift iteratively updates the mean-shift vector with the following formula:

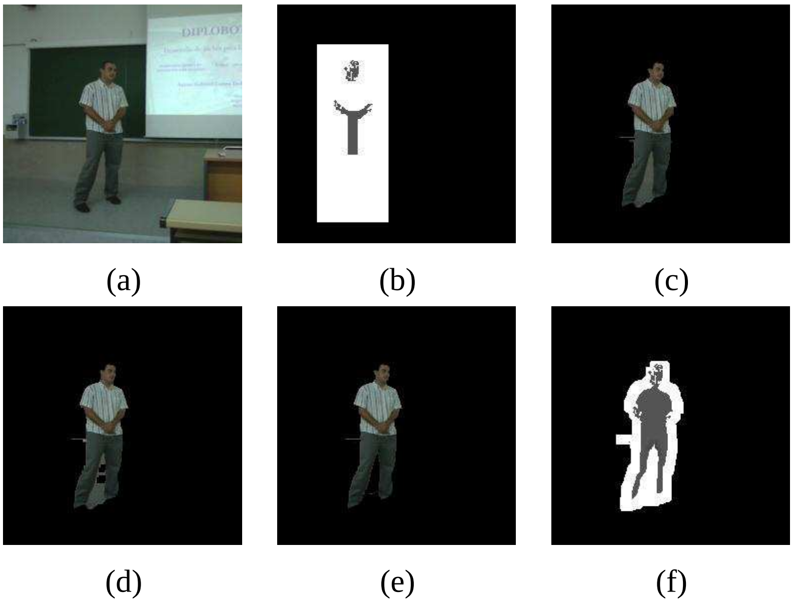
$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g(\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \|^2)}{\sum_{i=1}^{n} g(\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \|^2)} \quad (5)$$

until it converges, where $\mathbf{x}_i$ contains the value of pixel z_i in CIELuv space and its spatial coordinates, and returns the centers of the clusters (distribution modes) found. After convergence, we obtain a segmentation $\boldsymbol{\alpha}^t$ and the updated foreground and background GMMs $\boldsymbol{\theta}^t$ at frame f_t, which are used for further initialization at frame f_{t+1}. The result of this step is shown in Figure 2(c). Finally, we refine the segmentation of frame f_t eliminating false positive foreground pixels. By definition of the energy minimization scheme, GrabCut tends to find convex segmentation masks having a lower perimeter, given that each pixel on the boundary of the segmentation mask contributes on the global cost. Therefore, in order to eliminate these background pixels (commonly in concave regions) from the foreground segmentation, we re-initialize the trimap T as follows

$$\begin{aligned} T_B &= \{z_i|\alpha_i = 0\} \cup \\ &\left\{ z_i | \frac{\sum_{k=t-j}^{t} p(z_i|\alpha_i = 0, k_i, \boldsymbol{\theta}^k)}{j} > \frac{\sum_{k=t-j}^{t} p(z_i|\alpha_i = 1, k_i, \boldsymbol{\theta}^k)}{j} \right\} \\ T_F &= \{z_i \in \delta(z_i, h_{skin})\} \\ T_U &= \{z_i|\alpha_i = 1\} \setminus T_B \setminus T_F \end{aligned} \tag{6}$$

where the pixel background probability membership is computed using the GMM models of previous j segmentations. This formulation can also be extended to detect false negatives. However, in our case we focus on false positives since they appear frequently in the case of human segmentation. The result of this step is shown in Figure 2(d). Once the trimap has been redefined, false positive foreground pixels still remain, so the new set of seeds is used to iterate again GrabCut algorithm, resulting in a more accurate segmentation, as we can see in Figure 2(e).

Figure 2. STGrabcut pipeline example: (**a**) Original frame, (**b**) Seed initialization, (**c**) GrabCut, (**d**) Probabilistic re-assignment, (**e**) Refinement and (**f**) Initialization mask for f_{t+1}.



2.4. Temporal Extension

Considering A as the binary image representing $\boldsymbol{\alpha}$ at f_t (the one obtained before the refinement), we initialize the trimap for f_{t+1} as follows

$$\begin{aligned} T_F &= \{z_i \in I | z_i \in A \ominus ST_e, \alpha(z_i) = 1\} \\ T_U &= \{z_i \in I | z_i \in A \oplus ST_d, \alpha(z_i) = 1\} \setminus T_F \\ T_B &= \{z_i, z_i \in I\} \setminus (T_F \cup T_U) \end{aligned} \tag{7}$$

where $\ominus$ and $\oplus$ are erosion and dilation operations with their respective structuring elements ST_e and ST_d, $\alpha_i := \alpha(z_i)$, and $\setminus$ represents the set difference operation. The structuring elements are simple squares of a given size depending on the size of the person and the degree of movement we allow from f_t to f_{t+1}, assuming smoothness in the movement of the person. An example of a morphological mask is shown in Figure 2(f). Spatial information could be also included in the mean-shift algorithm in conjunction with color and spatial information. However, we included this information explicitly to be anisotropic. The whole segmentation methodology is detailed in the ST-GrabCut Algorithm 2.

Algorithm 2 Spatio-Temporal GrabCut algorithm.

```
1:  Person detection on $f_1$.
2:  Face detection and skin color model learning.
3:  Trimap $T$ initialization with detected bounding box and learnt skin color model.
4:  Initialize $\alpha_i = 0$ for $i \in T_B$ and $\alpha_i = 1$ for $i \in T_U \cup T_F$.
5:  Initialize Background and Foreground GMMs from sets $\alpha_i = 0$ and $\alpha_i = 1$ respectively, with
    Mean-shift.
6:  for $t$ = 1 ... $M$
7:      Person detection on $f_t$.
8:      Assign GMM components to pixels of $f_t$.
9:      Learn GMM parameters from data z.
10:     Estimate segmentation: Graph-cuts.
11:     Repeat from step 8, until convergence.
12:     Re-initialize trimap $T$ (Equation (6)).
13:     Assign GMM components to pixels.
14:     Learn GMM parameters from data z.
15:     Estimate segmentation: Graph-cuts.
16:     Repeat from step 12, until convergence.
17:     Initialize trimap $T$ using segmentation obtained in step 11 after convergence (equation 7) for $f_{t+1}$.
18: end for
```

2.5. Face Fitting

Once we have properly segmented the body region, the next step consists of fitting the face and the body limbs. For the case of face recovery, we base our procedure on mesh fitting using AAM, combining Active Shape Models and color and texture information [13].

AAM is generated by combining a model of shape and texture variation. First, a set of points are marked on the face of the training images that are aligned, and a statistical shape model is build [14]. Each training image is warped so the points match those of the mean shape. This is raster scanned into a texture vector, $\mathbf{g}$, which is normalized by applying a linear transformation, $\mathbf{g} \mapsto (\mathbf{g} - \mu_g \mathbf{1})/\sigma_g$, where $\mathbf{1}$ is a vector of ones, and μ_g and σ_g^2 are the mean and variance of elements of $\mathbf{g}$. After normalization, $\mathbf{g}^T \mathbf{1} = 0$ and $|\mathbf{g}| = 1$. Then, principal component analysis is applied to build a texture model. Finally, the correlations between shape and texture are learnt to generate a combined appearance model. The appearance model has parameter $\mathbf{c}$ controlling the shape and texture according to

$$x = \overline{x} + \mathbf{Q}_s \mathbf{c} \tag{8}$$

$$g = \overline{g} + \mathbf{Q}_g \mathbf{c} \tag{9}$$

where $\overline{x}$ is the mean shape, $\overline{g}$ the mean texture in a mean shaped patch, and $\mathbf{Q}_s$, $\mathbf{Q}_g$ are matrices designing the modes of variation derived from the training set. A shape $\mathbf{X}$ in the image frame can be generated by applying a suitable transformation to the points, $\mathbf{x} : \mathbf{X} = S_t(\mathbf{x})$. Typically, S_t will be a similarity transformation described by a scaling s, an in-plane rotation, θ, and a translation (t_x, t_y).

Once constructed the AAM, it is deformed on the image to detect and segment the face appearance as follows. During matching, we sample the pixels in the region of interest $\mathbf{g}_{im} = T_u(\mathbf{g}) = (u_1 + 1)\mathbf{g}_{im} + u_2 \mathbf{1}$, where $\mathbf{u}$ is the vector of transformation parameters, and project into the texture model frame, $\mathbf{g}_s = T_u^{-1}(\mathbf{g}_{im})$. The current model texture is given by $\mathbf{g}_m = \overline{g} + \mathbf{Q}_g \mathbf{c}$, and the difference between model and image (measured in the normalized texture frame) is as follows

$$\mathbf{r}(\mathbf{p}) = \mathbf{g}_s - \mathbf{g}_m \tag{10}$$

Given the error $E = |\mathbf{r}|^2$, we compute the predicted displacements $\delta \mathbf{p} = -\mathbf{R}\mathbf{r}(\mathbf{p})$, where $\mathbf{R} = \left(\frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \frac{\partial \mathbf{r}}{\partial \mathbf{p}}\right)^{-1} \frac{\partial \mathbf{r}^T}{\partial \mathbf{p}}$. The model parameters are updated $\mathbf{p} \mapsto \mathbf{p} + k\delta\mathbf{p}$, where initially $k = 1$. The new points $\mathbf{X}'$ and model frame texture $\mathbf{g}'_m$ are estimated, and the image is sampled at the new points to obtain $\mathbf{g}'_{mi}$ and the new error vector $\mathbf{r}' = T_{u'}^{-1}(g'_{im}) - g'_m$. A final condition guides the end of each iteration: if $|\mathbf{r}'|^2 < E$, then we accept the new estimate, otherwise, we set to $k = 0.5$, $k = 0.25$, and so on. The procedure is repeated until no improvement is made to the error.

With the purpose to discretize the head pose between frontal face and profile face, we create three AAM models corresponding to the frontal, right, and left view. Aligning every mesh of the model, we obtain the mean of the model. Finally, to determine the class of a fitted face by AAM models, that is given by its proximity to the closest mean model.

Taking into account the discontinuity that appears when a face moves from frontal to profile view, we use three different AAM corresponding to three meshes of 21 points: frontal view $\Im_F$, right lateral view $\Im_R$, and left lateral view $\Im_L$. In order to include temporal and spatial coherence, meshes at frame f_{t+1}

are initialized by the fitted mesh points at frame f_t. Additionally, we include a temporal change-mesh control procedure, as follows

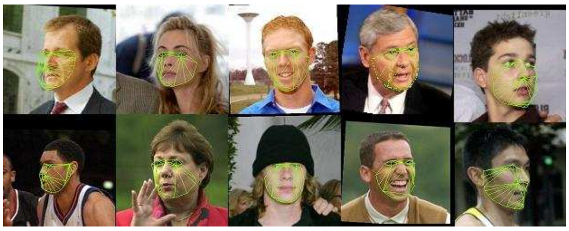
$$\Im^{t+1} = \min_{\Im^{t+1}}\{E_{\Im_F}, E_{\Im_R}, E_{\Im_L}\}, \Im^{t+1} \in \nu(\Im^t) \tag{11}$$

where $\nu(\Im^t)$ corresponds to the meshes contiguous to the mesh $\Im^t$ fitted at time t (including the same mesh), and $E_{\Im_i}$ is the fitting error cost of mesh $\Im_i$. This constraint avoids false jumps and imposes smoothness in the temporal face behavior (e.g., a jump from right to left profile view is not allowed).

In order to obtain more accurate pose estimation, after fitting the mesh, we take advantage of its variability to differentiate among a set of head poses. Analyzing the spatial configuration of the 21 landmarks that composes a mesh, we create a new training set divided in five classes. We define five different head poses as follows: right, middle-right, frontal, middle-left, and left. In the training process, every mesh has been aligned, and PCA is applied to save the 20 most representative eigenvectors. Then, a new image is projected to that new space and classified to one of the five different head poses according to a 3-Nearest Neighbor rule.

Figure 3 shows examples of the AAM model fitting and pose estimation in images (obtained from [15]) for the five different head poses.

Figure 3. From left to right: left, middle-left, frontal, middle-right and right mesh fitting.



2.6. Pose Recovery

Considering the refined segmented body region obtained using the proposed ST-GrabCut algorithm, we construct a pictorial structure model [16]. We use the method of Ramanan [6,8], which captures the appearance and spatial configuration of body parts. A person's body parts are tied together in a tree-structured conditional random field. Parts, l_i, are oriented patches of fixed size, and their position is parameterized by location (x, y) and orientation ϕ. The posterior of a configuration of parts $L = l_i$ given a frame f_t is

$$P(L|f_t) \propto \exp\left(\sum_{(i,j)\in E} \Psi(l_i, l_j) + \sum_i \Phi(l_i|f_t)\right) \tag{12}$$

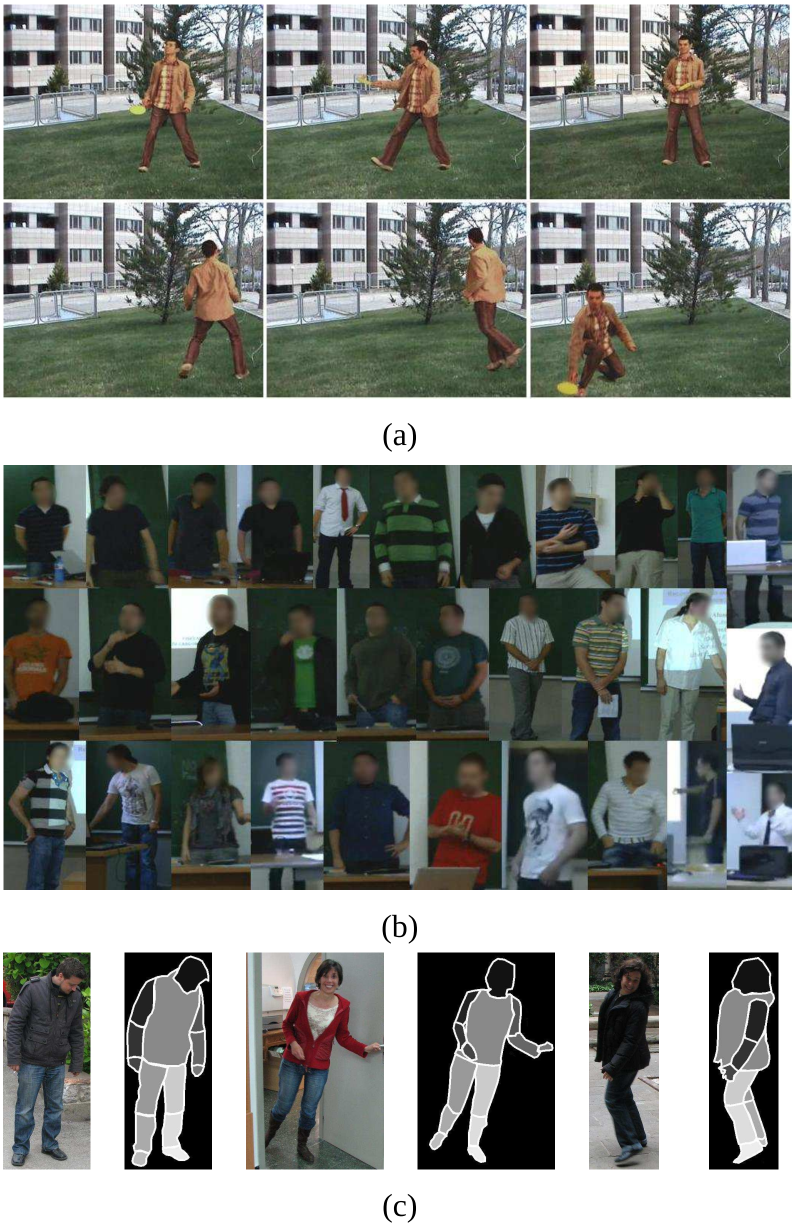
The pair-wise potential $\Psi(l_i, l_j)$ corresponds to a spatial prior on the relative position of parts and embeds the kinematic constraints. The unary potential $\Phi(l_i|I)$ corresponds to the local image evidence for a part in a particular position. Inference is performed over tree-structured conditional random field.

Since the appearance of the parts is initially unknown, a first inference uses only edge features in Φ. This delivers soft estimates of body part positions, which are used to build appearance models of the parts and background (color histograms). Inference is then repeated with Φ using both edges and appearance. This parsing technique simultaneously estimates pose and appearance of parts. For each body part, parsing delivers a posterior marginal distribution over location and orientation (x, y, ϕ) [6,8].

3. Results

Before the presentation of the results, we discuss the data, methods and parameters of the comparative, and validation measurements.

Figure 4. (**a**) Samples of the cVSG corpus and (**b**) UBDataset image sequences, and (**c**) HumanLimb dataset.


(a)

(b)

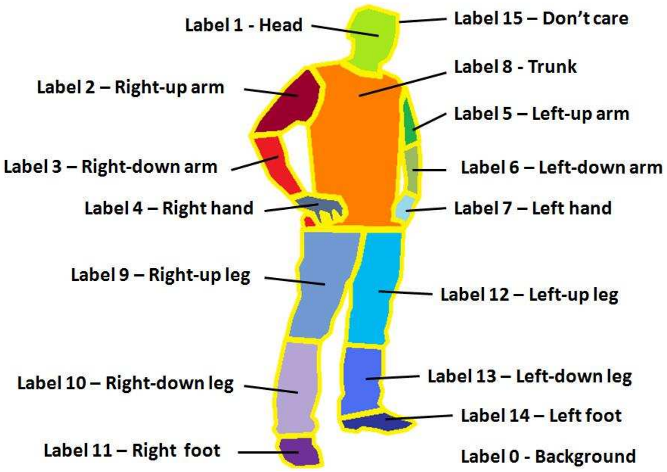
(c)

○ *Data*: We use the public image sequences of the Chroma Video Segmentation Ground Truth (cVSG) [17], a corpus of video sequences and segmentation masks of people. Chroma based techniques have been used to record Foregrounds and Backgrounds separately, being later combined to achieve final video sequences and accurate segmentation masks almost automatically. Some samples of the sequence we have used for testing are shown in Figure 4(a). The sequence has a total of 307 frames. This image sequence includes several critical factors that make segmentation difficult: object textural complexity, object structure, uncovered extent, object size, Foreground and Background velocity, shadows, background textural complexity, Background multimodality, and small camera motion.

As a second database, we have also used a set of 30 videos corresponding to the defense of undergraduate thesis at the University of Barcelona to test the methodology in a different environment (UBDataset). Some samples of this dataset are shown in Figure 4(b).

Moreover, we present the Human Limb dataset, a new dataset composed by 227 images from 25 different people. At each image, 14 different limbs are labeled (see Figure 4(c)), including the "do not care" label between adjacent limbs, as described in Figure 5. Backgrounds are from different real environments with different visual complexity. This dataset is useful for human segmentation, limb detection, and pose recovery purposes [18].

Figure 5. Human Limb dataset labels description.



○ *Methods*: We test the classical semi-automatic GrabCut algorithm for human segmentation comparing with the proposed ST-GrabCut algorithm. In the case of GrabCut, we set the number of GMM components $k = 5$ for both foreground and background models. Furthermore, the already trained models used for person and face detectors have been taken from the OpenCV 2.1.

We also test the mesh fitting and body pose recovery methodologies on the obtained segmentations. The body model used for the pose recovery was taken directly from the work of [8].

○ *Validation measurements*: In order to evaluate the robustness of the methodology for human body segmentation, face and pose fitting, we use the ground truth masks of the images to compute the overlapping factor O as follows

$$O = \frac{\sum M_{GC} \cap M_{GT}}{\sum M_{GC} \cup M_{GT}} \tag{13}$$

where M_{GC} and M_{GT} are the binary masks obtained for spatio-temporal GrabCut segmentation and the ground truth mask, respectively.

3.1. Spatio-Tempral GrabCut Segmentation

First, we test the proposed ST-GrabCut segmentation on the sequence from the public cVSG corpus. The results for the different experiments are shown in Table 1. In order to avoid the manual initialization of classical GrabCut algorithm, for all the experiments, seed initialization is performed applying the commented person HOG detection, face detection, and skin color model. First row of Table 1 shows the overlapping performance of Equation (13) applying GrabCut segmentation with k-means clustering to design the GMM models. Second row shows the overlapping performance considering the spatial extension of the algorithm introduced by using Mean Shift clustering (Equation (5)) to design the GMM models. One can see a slight improvement when using the second strategy. This is mainly because Mean Shift clustering takes into account spatial information of pixels in clustering time, which better defines contiguous pixels of image to belong to GMM models of foreground and background. Third performance in Table 1 shows the overlapping results adding the temporal extension to the spatial one, considering the morphology refinement based on previous segmentation (Equation (7)). In this case, we obtain near 10% of performance improvement respect the previous result. Finally, last result of Table 1 shows the full-automatic ST-GrabCut segmentation overlapping performance taking into account spatio-temporal coherence, and the segmentation refinement introduced in Equation (6). One can see that it achieves about 25% of performance improvement in relation with the previous best performance. Some segmentation results obtained by the GrabCut algorithm for the cVSG corpus are shown in Figure 6. Note that the ST-GrabCut segmentation is able to robustly segment convex regions. We have also applied the ST-GrabCut segmentation methodology on the image sequences of UBDataset. Some segmentations are shown in Figure 6.

Table 1. GrabCut and ST-GrabCut Segmentation results on cVSG corpus.

Approach	Mean overlapping
GrabCut	0.5356
Spatial extension	0.5424
Temporal extension	0.6229
ST-GrabCut	0.8747

3.2. Face Fitting

In order to measure the robustness of the spatio-temporal AAM mesh fitting methodology, we performed the overlapping analysis of meshes in both un-segmented and segmented image sequence of the public cVSG corpus. Overlapping results are shown in Table 2. One can see that the mesh fitting works fine in unsegmented images, obtaining a final mean overlapping of 89.60%. In this test, we apply HaarCascade face detection implemented and trained by the Open Source Computer Vision library (OpenCv). The face detection method implemented in OpenCV by Rainer Lienhart is very similar to the one published and patented by Paul Viola and Michael Jones, namely called Viola–Jones face detection method [19]. The classifier is trained with a few hundreds of sample views of a frontal face, that are scaled to the same size (20 × 20), and negative examples of the same size. However, note that combining the temporal information of previous fitting and the ST-GrabCut segmentation, the face mesh fitting considerably improves, obtaining a final of 96.36% of overlapping performance. Some example of face fitting using the AAM meshes for different face poses of the cVSG corpus are shown in Figure 7.

Figure 6. Segmentation examples of (**a**) UBDataset sequence 1, (**b**) UBDataset sequence 2 and (**c**) cVSG sequence.

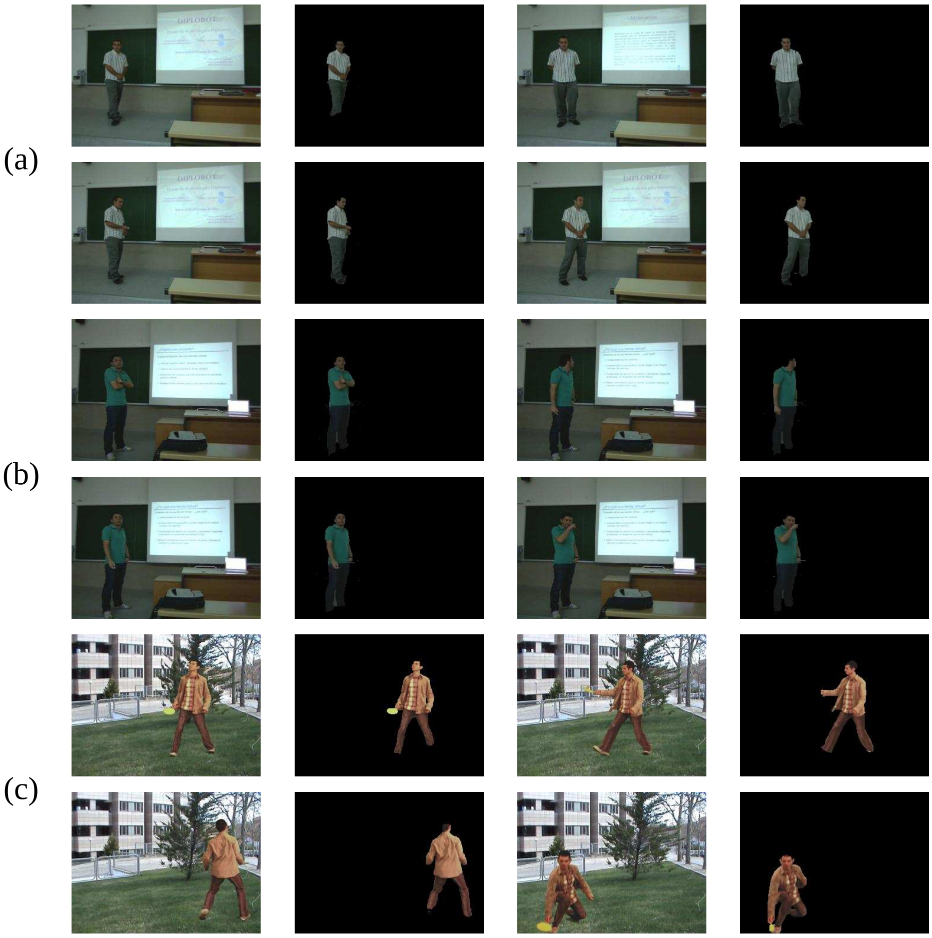
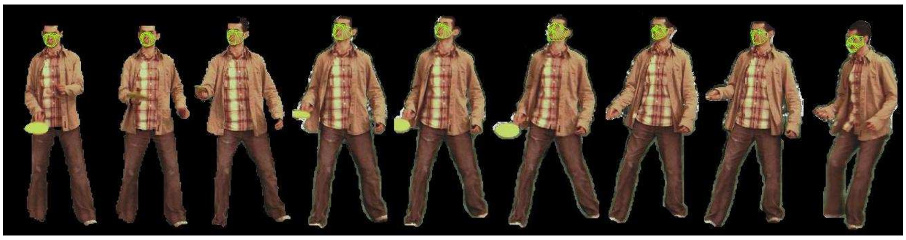


Figure 7. Samples of the segmented cVSG corpus image sequences fitting the different AAM meshes.



To create three AAM models that represent frontal, right and left views, we have created a training set composed by 1,000 images for each view. The images have been extracted from the public database [15]. To build three models we manually put 21 landmarks over 500 images for each view. The landmarks of the remaining 500 images which covers one view, has been placed by a semi-automatic process, applying

AAM with the set learnt and manually correcting. Finally, we align every resulting mesh and we obtain the mean for each model. As the head pose classifier, to classify the spatial mesh configuration in 5 head poses, we have labeled manually the class of the mesh obtained applying the closest AAM model. Every spatial mesh configuration is represented by the 20 most representative eigenvectors. The training set is formed by 5,000 images from the public database [15]. Finally, we have tested the classification of the five face poses on the cVSG corpus, obtaining the percentage of frames of the subject at each pose. The obtained percentages are shown in Table 3.

Table 2. AAM mesh fitting on original images and segmented images of the cVSG corpus.

Approach	Mean overlapping
Mesh fitting without segmentation	0.8960
ST-Grabcut & Temporal mesh fitting	0.9636

Table 3. Face pose percentages on the cVSG corpus.

Face view	System classification	Real classification
Left view	0.1300	0.1211
Near Left view	0.1470	0.1347
Frontal view	0.2940	0.3037
Near Right view	0.1650	0.1813
Right view	0.2340	0.2590

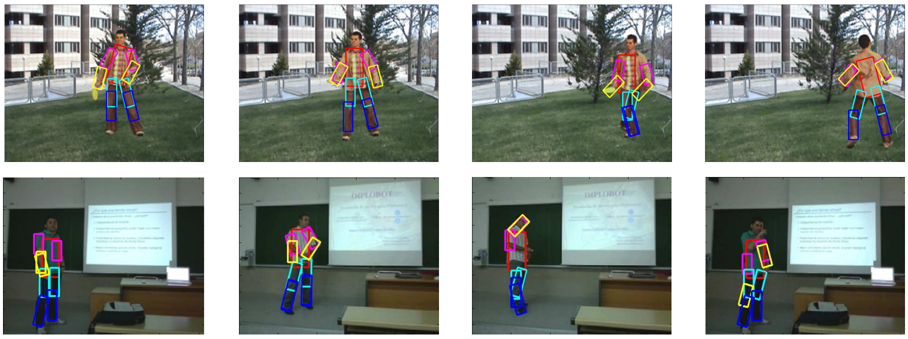
3.3. Body Limbs Recovery

Finally, we combine the previous segmentation and face fitting with a full body pose recovery [8]. In order to show the benefit of applying previous ST-GrabCut segmentation, we perform the overlapping performance of full pose recovery with and without human segmentation, always within the bounding box obtained from HOG person detection. Results are shown in Table 4. One can see that pose recovery considerably increases its performance when reducing the region of search based on ST-GrabCut segmentation. Some examples of pose recovery within the human segmentation regions for cVSG corpus and UBdataset are shown in Figure 8. One can see that in most of the cases body limbs are correctly detected. Only in some situations, occlusions or changes in body appearance can produce a wrong limb fitting.

Table 4. Overlapping of body limbs based on ground truth masks.

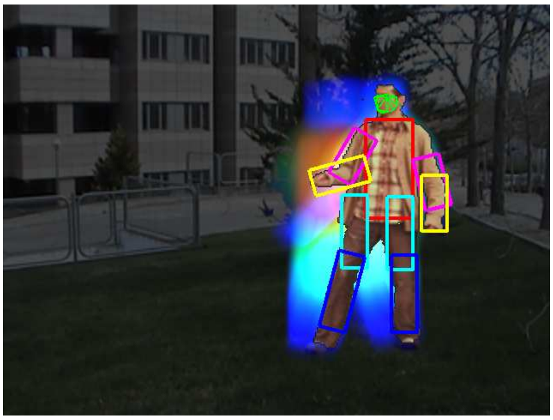
Approach	Mean overlapping
Limb recovery without segmentation	0.7919
ST-Grabcut & Limb recovery	0.8760

Figure 8. Pose recovery results in cVSG sequence.



In Figure 9 we show the application of the whole framework to perform temporal tracking, segmentation and full face and pose recovery. The colors correspond to the body limbs. The colors increase in intensity based on the instant of time of its detection. One can see the robust detection and temporal coherence based on the smooth displacement of face and limb detections.

Figure 9. Application of the whole framework (pose and face recovery) on an image sequence.



3.4. Human Limb Data Set

In this last experiment, we test our methodology on the presented Human Limb dataset. From the 14 total limb annotations, we grouped them into six categories: trunk, up-arms, up-legs, low-arms, low-legs, and head, and we tested the full pose recovery framework. In this case, we tested the body limb recovery with and without applying the ST-GrabCut segmentation, and computed three different overlapping measures: (1) %, which corresponds to the overlapping percentage defined in Equation (13); (2) wins, which corresponds to the number of Limb regions with higher overlapping comparing both strategies; (3) match, which corresponds to the number of limb recoveries with overlapping superior to 0.6. The results are shown in Table 5. One can see that because of the reduced region where the subjects appear, in most cases there is no significant difference applying the limb recovery procedure with or without previous segmentation. Moreover, the segmentation algorithm is not working at maximum performance due to the same reason, since very small background regions are present in the images, and thus the background color model is quite poor. Furthermore, in this dataset we are working with images, not videos, and for this reason we cannot include the temporal extension in our ST-GrabCut algotithm for this experiment. On the other hand, looking at the mean average overlapping in the last column of the table, one can see that ST-GrabCut improves for all overlapping measures the final limb overlapping. In particular, in the case of the Low-legs recovery is when a more clear improvement appears using ST-GrabCut segmentation. The part of the image corresponding to Low-legs is where more background influence exists, and thus the limb recovery has the highest confusion. However, as ST-GrabCut is able to properly segment the concave regions of the Low-legs regions, a significant improvement is obtained when applying the limb recovery methodology. Some results are illustrated on the images of Figure 10, where the images on the bottom correspond to the improvements obtained using the ST-GrabCut algorithm. Finally, Figure 11 show examples of the face fitting methodology applied on the human body limb dataset.

Table 5. Overlapping percentages between body parts (intersection over union), wins (comparing the highest overlapping with and without segmentation), and matching (considering only overlapping greater than 0.6).

		Trunk	**Up-arms**	**Up-legs**	**Low-arms**	**Low-legs**	**Head**	**Mean**
%	**No segmentation**	0.58	0.53	0.59	0.50	0.48	0.67	0.56
	STGrabCut*	0.58	0.53	0.58	0.50	0.56	0.67	**0.57**
Wins	**No segmentation**	106	104	108	109	68	120	102.5
	STGrabCut*	121	123	119	118	159	107	**124.5**
Match	**No segmentation**	133	127	130	121	108	155	129
	STGrabCut*	125	125	128	117	126	157	**129.66**

* STGrabCut was used without taking into account temporal information.

Figure 10. Human Limb dataset results. Up row: limb recovery without ST-GrabCut segmentation. Down row: limb recovery with ST-GrabCut segmentation.

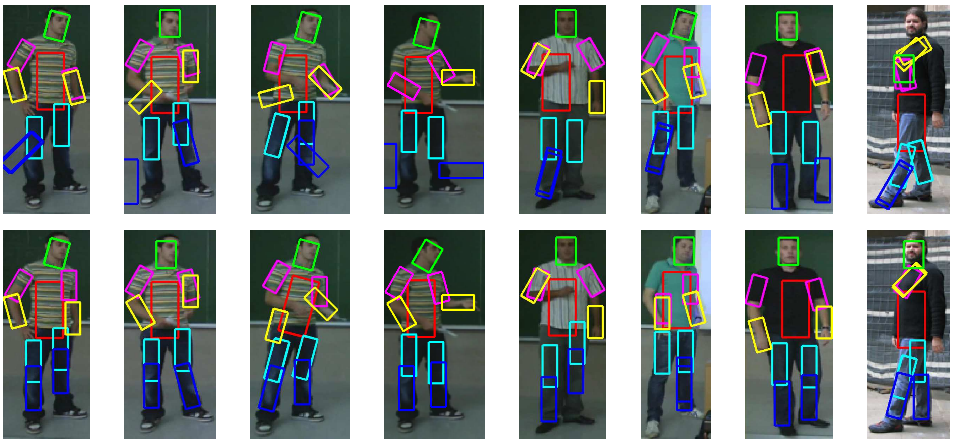
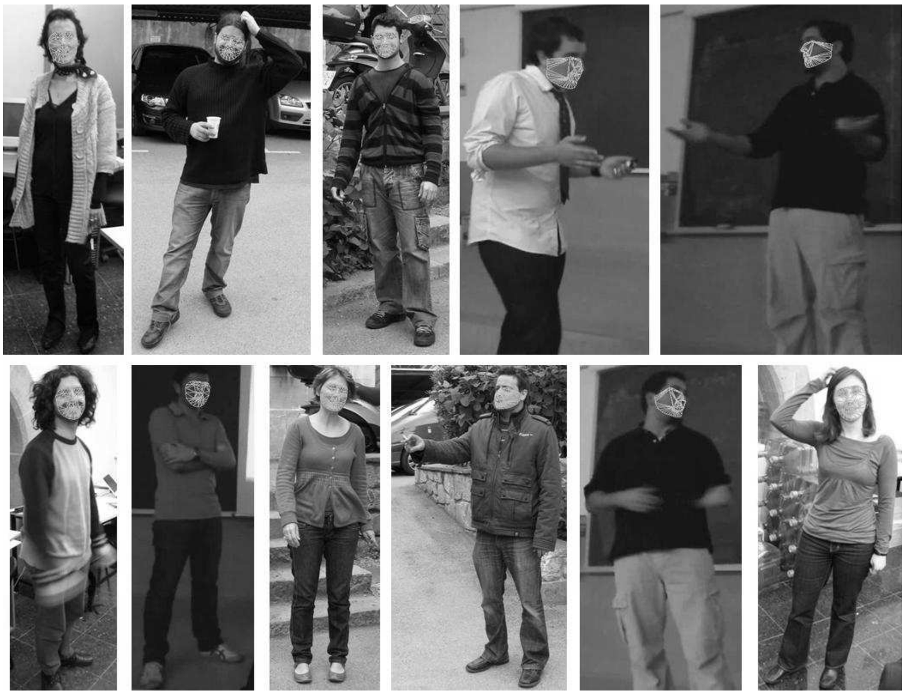


Figure 11. Application of face recovery on human body limb dataset.



4. Conclusions

In this paper, we presented an evolution of the semi-automatic GrabCut algorithm for dealing with the problem of human segmentation in image sequences. The new full-automatic ST-GrabCut algorithm uses a HOG-based person detector, face detection, and skin color model to initialize GrabCut seeds. Spatial coherence is introduced via Mean Shift clustering, and temporal coherence is considered based on the historical of Gaussian Mixture Models. The segmentation procedure is combined with Shape and Active Appearance models to perform full face and pose recovery.

This general and full-automatic human segmentation, pose recovery, and tracking methodology showed higher performance than classical approaches in public image sequences and a novel Human Limb dataset from uncontrolled environments, which makes it useful for general human face and gesture analysis applications.

One of the limitations of the method is that it depends on the initialization of the ST-GrabCut algorithm, which basically depends on the person and face detectors. Initially, we wait until at least one bounding box is returned by the person detector. This is a critical point, since we will trust the first detection and start segmenting with this hypothesis. In contrast, there is no problem if a further detection is missed, since we initialize the mask with the previous detection (temporal extension). Moreover, due to its sequential application, false seed labeling can accumulate segmentation errors along the video sequence. As the next step, we plan to extend the limb recovery approach so that more complex poses and gestures can be recognized, and feed a gesture recognition system [20] with the temporal aggregation of the recovered poses along the sequence in order to look for motion patterns of the limbs.

As a future work, the algorithm could be extended in order to segment sequences with more than one person present in the images, since our current method only segments one subject in the scene.

Acknowledgements

This work has been supported in part by projects IMSERSO-Ministerio de Sanidad 2011 Ref. MEDIMINDER, RECERCAIXA 2011 Ref. REMEDI, TIN2009-14404-C02 and CONSOLIDER-INGENIO CSD 2007-00018. The work of Antonio is supported by an FPU fellowship from the Spanish government.

References

1. Dalal, N.; Triggs, B. Histogram of Oriented Gradients for Human Detection. In Proceedings of CVPR ’05: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 25 June 2005; Volume 2, pp. 886–893.
2. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154.
3. Geronimo, D.; Lopez, A.; Sappa, A. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. *IEEE Trans. Patt. Anal. Mach. Intell.* **2010**, *32*, 1239–1258.
4. Rother, C.; Kolmogorov, V.; Blake, A. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314.
5. Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Progressive Search Space Reduction for Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 24–26 June 2008.
6. Ferrari, V.; Marin, M.; Zisserman, A. Pose Search: Retrieving People Using Their Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 20–25 June 2009.
7. Corrigan, D.; Robinson, S.; Kokaram, A. Video Matting Using Motion Extended GrabCut. In *Proceedings of 5th IET European Conference on Visual Media Production (CVMP)*, London, UK, 26–27 November 2008.
8. Ramanan, D. Learning to Parse Images of Articulated Bodies. NIPS, 2006. Available online: http://books.nips.cc/papers/files/nips19/NIPS2006_0899.pdf (accessed on 8 November 2012).
9. Boykov, Y.Y.; Jolly, M.P. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images. In *Proceedings of ICCV 2001: Eighth IEEE International Conference on Computer Vision*, Vancouver, BC, Canada, 7–14 July 2001.
10. Boykov, Y.; Funka-Lea, G. Graph Cuts and Efficient N-D Image Segmentation. *Int. J. Comput. Vis.* **2006**, *70*, 109–131.
11. Kolmogorov, V.; Zabih, R. What Energy Functions can be Minimized via Graph Cuts. *IEEE Trans. Patt. Anal. Mach. Intell.* **2004**, *26*, 65–81.
12. Boykov, Y.; Kolmogorov, V. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. Patt. Anal. Mach. Intell.* **2001**, *26*, 359–374.
13. Cootes, T.; Edwards, J.; Taylor, C. Active Appearance Models. *IEEE Trans. Patt. Anal. Mach. Intell.* **2001**, *23*, 681–685.
14. Cootes, T.; Taylor, C.; Cooper, D.; Graham, J. Active Shape Models—Their Training and Application. *Comput. Vis. Image Understand.* **1995**, *61*, 38–59.

15. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report 07-492007; University of Massachusetts: Amherst, MA, USA, 2007.
16. Felzenszwalb, P.; Huttenlocher, D. Pictorial Structures for Object Recognition. *Int. J. Comput. Vis.* **2005**, *61*, 55–79.
17. Tiburzi, F.; Escudero, M.; Bescos, J.; Martinez, J. A Ground-Truth for Motion-Based Video-Object Segmentation. In *Proceedings of IEEE International Conference on Image Processing (Workshop on Multimedia Information Retrieval)*, San Diego, CA, USA, 12–15 October 2008.
18. Human Limb dataset. Availbel online: http://www.maia.ub.es/%7Esergio/linked/humanlimbdb.zip (accessed on 8 November 2012).
19. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Inte. J. Comput. Vision* **2004**, *57*, 137–154.
20. Alon, J.; Athitsos, V.; Yuan, Q.; Sclaroff, S. A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 1685–1699.