Spatio-Temporal GrabCut Human Segmentation for Face and Pose Recovery

Antonio Hernández[1] ahernandez@cvc.uab.es
Miguel Reyes[1] mreyese@gmail.com
Sergio Escalera[2] sergio@maia.ub.es
Petia Radeva[1] petia@cvc.uab.es

[1] Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain.
[2] Dept. Matemàtica Aplicada i Anàlisi, University of Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain.

Abstract

In this paper, we present a full-automatic Spatio-Temporal GrabCut human segmentation methodology. GrabCut initialization is performed by a HOG-based subject detection, face detection, and skin color model for seed initialization. Spatial information is included by means of Mean Shift clustering whereas temporal coherence is considered by the historical of Gaussian Mixture Models. Moreover, human segmentation is combined with Shape and Active Appearance Models to perform full face and pose recovery. Results over public data sets as well as proper human action base show a robust segmentation and recovery of both face and pose using the presented methodology.

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, or 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 perform over a reduced set of small candidate regions. In [5], the authors propose a full-body detector based on a cascade of classifiers [13] using HOG features. This methodology is currently being used in several works related to the pedestrian detection problem [8]. GrabCut [11] 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 [7]. In the case of working with sequences of images, this optimization problem can also be considered to have temporal coherence. In the work of [4], 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 determined a region of interest, 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 [10], 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 scheme and optimized using believe 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 [7].

In this paper, we propose a full-automatic Spatio-Temporal GrabCut human segmentation methodology. 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 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 CRF model of [10] is applied over the obtained segmentation, showing high robustness capturing body limbs because of the accurate human segmentation.

The rest of the paper is organized as follows: Section 2 describes the proposed methodology, presenting the spatio-temporal GrabCut segmentation, the Active Appearance models for face fitting, and the pose recovery methodology. Experimental results on two different data sets 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 Shape and Active Appearance Models used to recover the face, and the body pose recovery methodology based on the approach of [10]. All methods presented in this section are combined to improve final segmentation and pose recovery. Figure 1 illustrates the different modules of the project.

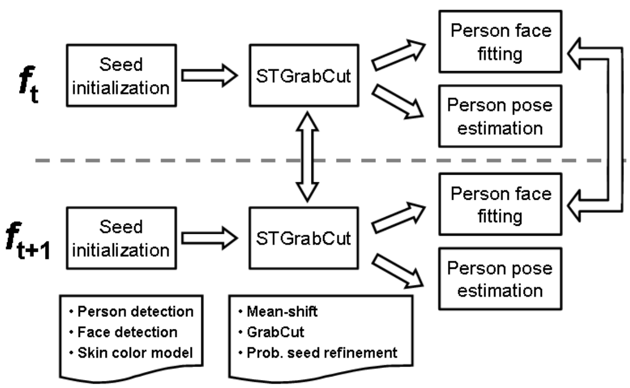


Figure 1. Overall block diagram of the methodolody

2.1. Spatio-Temporal GrabCut segmentation

In [11], the authors proposed an approach to find a binary segmentation -Background, Foreground- of an image by formulating an energy minimization scheme as the one presented in [1], extended using color instead of just gray-scale information. Given a color image, 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 on 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 -that 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

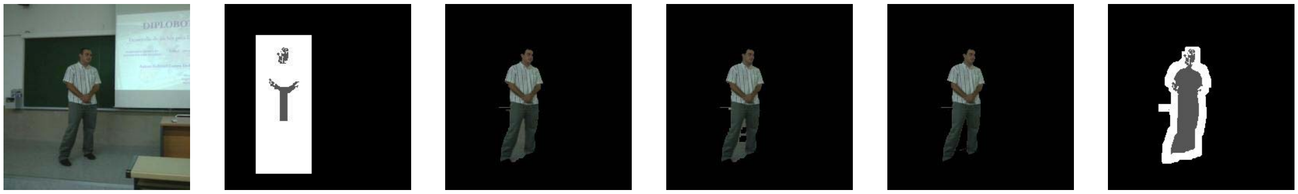
$$\mathbf{V}(\boldsymbol{\alpha}, \mathbf{z}) = \gamma \sum_{\{m,n\} \in C} [\alpha_n \neq \alpha_m] \exp\left(-\beta \|z_m - z_n\|^2\right) \tag{4}$$

With this energy minimization scheme and given the initial trimap T, the final segmentation is performed using a minimum cut algorithm [1]. The classical semi-automatic GrabCut algorithm is summarized in Algorithm 2.1.

Algorithm 2.1: Original GrabCut algorithm.

1: Trimap T initialization with manual annotation.
2: Initialize $a_i = 0$ for $n \in T_B$ and $a_i = 1$ for $n \in T_U \cup T_F$.
3: Initialize Background and Foreground GMMs from sets $a_n = 0$ and $a_n = 1$ 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

Our proposal bases on the previous GrabCut framework, focusing on human body segmentation 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 [5]. 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



(a) (b) (c) (d) (e) (f)

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}

accuracy of the segmentation algorithm, we include Foreground seeds exploiting spatial and appearance prior information. On one hand, we define a small central region R inside B and set these pixels as Foreground. On the other, we apply a face detector based on a cascade of classifiers using Haar-like features [13] over B, and learn a skin color model h_{skin}. 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).

Once we have initialized the trimap, we apply the iterative minimization algorithm shown in steps 4 to 7 of original GrabCut (algorithm 2.1). However, instead of applying k-means for the initialization of the GMMs we use Mean-Shift clustering, which also takes into account spatial coherence. 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

$$\begin{aligned} T_B &= \{z_i | \alpha_i = 0\} \cup \\ & \left\{ z_i | \sum_{k=t-j}^{t} \frac{p(z_i|\alpha_i = 0, k_i, \boldsymbol{\theta}^k)}{j} > \sum_{k=t-j}^{t} \frac{p(z_i|\alpha_i = 1, k_i, \boldsymbol{\theta}^k)}{j} \right\} \\ T_U &= \{z_i | \alpha_i = 1\} \setminus T_B \\ T_F &= \{z_i \in \delta(z_i, h_{skin})\} \end{aligned} \tag{5}$$

where the pixel background probability membership is computed using the GMM models of previous segmentations. 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 Fig. 2(e). Finally, 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

$$\begin{aligned} T_F &= \{z_i \in A \ominus ST_d\} \\ T_U &= \{z_i \in A \oplus ST_e\} \setminus T_F \\ T_B &= \{z_i, i = 1..N\} \setminus (T_F \cup T_U) \end{aligned} \tag{6}$$

where $\ominus$ and $\oplus$ are erosion and dilation operations with their respective structuring elements ST_d and ST_e. 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). The whole segmentation methodology is detailed in the ST-GrabCut algorithm 2.2.

Algorithm 2.2: Proposed ST-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 $a_i = 0$ for $n \in T_B$ and $a_i = 1$ for $n \in T_U \cup T_F$.
5: Initialize Background and Foreground GMMs from sets $a_n = 0$ and $a_n = 1$ 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 5).
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 6) for f_{t+1}.
18: **end for**

2.2. Face fitting

Once we have properly segmented the body region, 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 Active Appearance Models (AAM), that benefits from Active Shape Models and color and texture information [2].

Active Appearance Model 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 [3]. 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, eigenanalysis 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

$$x = \overline{x} + \mathbf{Q}_s\mathbf{c} \quad (7)$$

$$g = \overline{g} + \mathbf{Q}_g\mathbf{c} \quad (8)$$

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 $\mathbf{g}_m = \overline{g} + \mathbf{Q}_g\mathbf{c}$. The current difference between model and image (measured in the normalized texture frame) is as follows

$$\mathbf{r}(\mathbf{p}) = \mathbf{g}_s - \mathbf{g}_m \quad (9)$$

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 $\mathbf{g}'_{mi}$, obtaining the new error vectoras $\mathbf{r}' = T_{u'}^{-1}(g'_{im}) - g'_m$. 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.

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 laterial view $\Im_R$, and left laterial 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

$$\Im^{t+1} = \min_{\Im^{t+1}}\{E_{\Im_F}, E_{\Im_R}, E_{\Im_L}\}, \Im^{t+1} \in \nu(\Im^t) \quad (10)$$

where $\nu(\Im^t)$ corresponds to the meshes contiguous to the mesh $\Im^t$ fitted at time t (including the same mesh). 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 a more accurate pose estimation, after fitting the mesh, we take advantage of its variability to differentiate among a set of head poses. We have defined five different head poses: right, middle-right, frontal, middle-left, and left. In order to define this set, the fitted frontal meshes in the training set are classified in three different poses: middle-right, frontal, and middle left, whereas the training samples of the left and right meshes are directly classified in full-left and full-right poses, respectively. In order to learn the five different head poses, training images are aligned, and PCA is applied to save the 20 most representative eigenvectors. Then, a new mesh 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 in natural images (obtained from [9]) for the three different meshes.

Figure 3. From left to right: left, frontal, and right mesh fitting using AAM.

2.3. Pose recovery

Considering the refined segmented body region, we construct a pictorial structure model [6]. We use the method of Ramanan [10, 7], 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

$$P(L|f_t) \propto \exp\left(\sum_{(i,j)\in E} \Psi(l_i, l_j) + \sum_i \Phi(l_i|f_t)\right) \quad (11)$$

The pairwise 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 by sum-product Belief Propagation.

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 in 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) [10, 7].

(a)

(b)

Figure 4. (a) Samples of the cVSG corpus and (b) UBDataset image sequences.

3. Results

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

∘ *Data*: We use the public image sequences of the Chroma Video Segmentation Ground Truth (cVSG) [12], 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. Alternatively 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 data set are shown in Figure 4(b).

∘ *Methods*:

We test the classical semi-automatic GrabCut algorithm for human segmentation comparing with the proposed ST-GrabCut algorithm. We also test the mesh fitting and body pose recovery methodologies over the obtained segmentation.

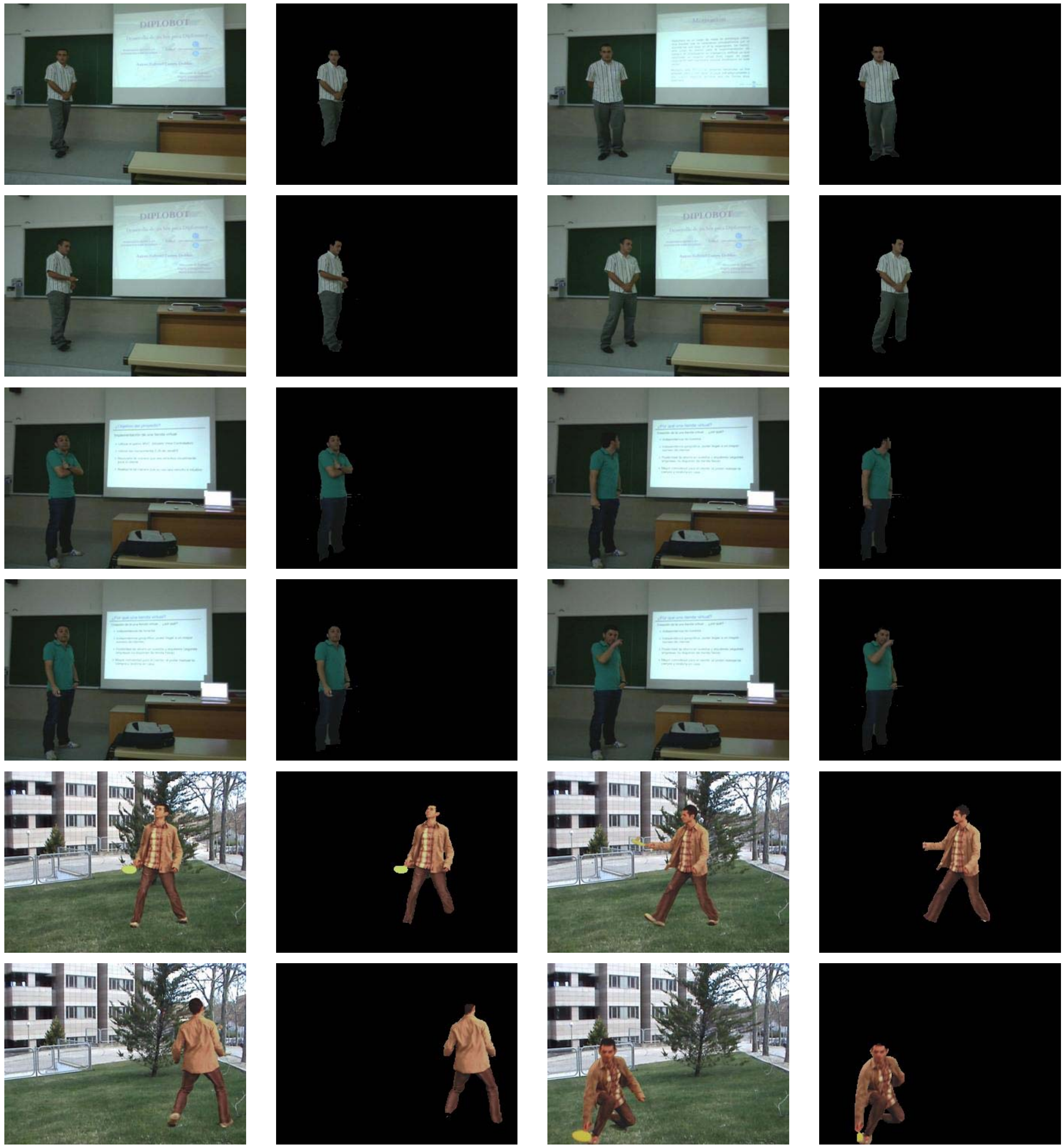
∘ *Validation measurements*: In order to evaluate the robustness of the methodology for human body segmentation, face and pose fitting, we use the ground truth masks of the video sequences to compute the overlapping factor O as follows

$$O = \frac{\sum M_{GC} \cap M_{GT}}{\sum M_{GC} \cup M_{GT}} \quad (12)$$

where M_{GC} and M_{GC} 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 12 applying GrabCut segmentation with k-means clustering to design the GMM models. Second row shows the overlapping performance considering Mean Shift clustering to design the GMM models. One can see a slight improvement when using the second strategy. This is mainly due to the fact that 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 considering the morphology refinement based



(a)

(b)

(c)

Figure 5. Segmentation examples of (a) UBDataset sequence 1, (b) UBDataset sequence 2 and (c) cVSG sequence.

on previous segmentation. 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. 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 5. 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 5.

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. Over-

Figure 6. Samples of the segmented cVSG corpus image sequences fitting the different AAM meshes.

Approach	Mean overlapping
K-means	0.5356
Mean-shift	0.5424
Morphology	0.6229
ST-GrabCut	0.8747

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

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

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

Face view	Percentage of frames
Left view	0.1300
Near Left view	0.1470
Frontal view	0.2940
Near Right view	0.1650
Right view	0.2340

Table 3. Face pose percentages on the cVSG corpus.

lapping results are shown in Table 3. One can see that the mesh fitting works fine in unsegmented images, obtaining a final mean overlapping of 89.60%. 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 6.

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.

3.3. Body limbs recovery

Finally, we combine the previous segmentation and face fitting with a full body pose recovery [10]. 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

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

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

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 7. 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.

Finally, in Figure 8 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 base on the smooth displacement of face and limb detections.

4. Conclusion

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 from uncontrolled environments, which makes it useful for general human face and gesture analysis applications.

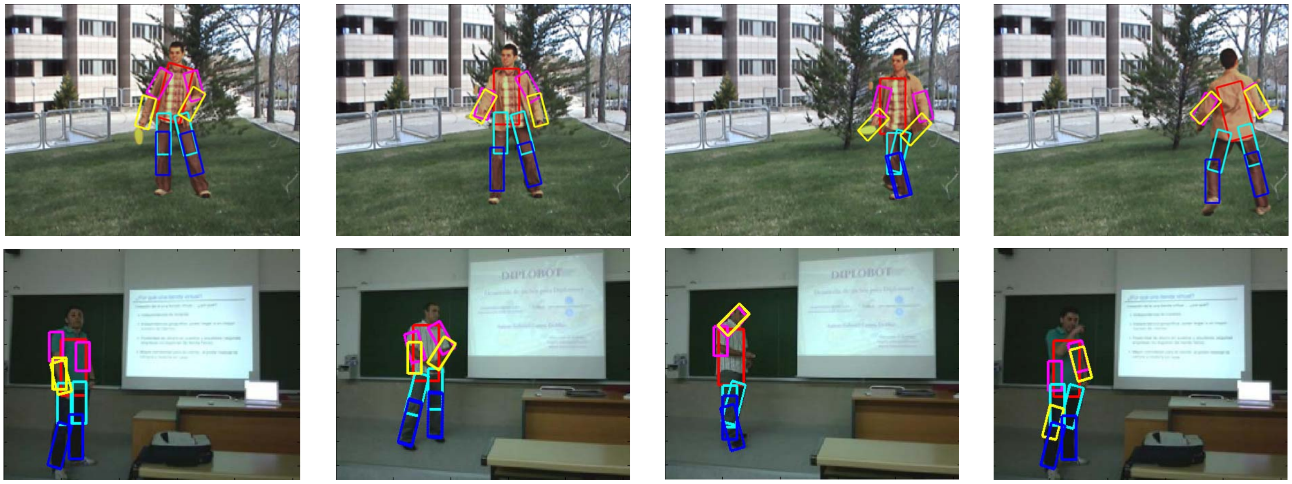


Figure 7. Pose recovery results in cVSG sequence

One of the limitations of the method is that it depends on the initialization of the ST-GrabCut algorithm. Moreover, due to its sequential application, false seed labeling can accumulate segmentation errors along the video sequence. As next step we plan to extend the limb recovery approach so that more complex poses and gestures can be recognized.

Acknowledgements

This work has been supported in part by projects TIN2009-14404-C02 and CONSOLIDER-INGENIO CSD 2007-00018.

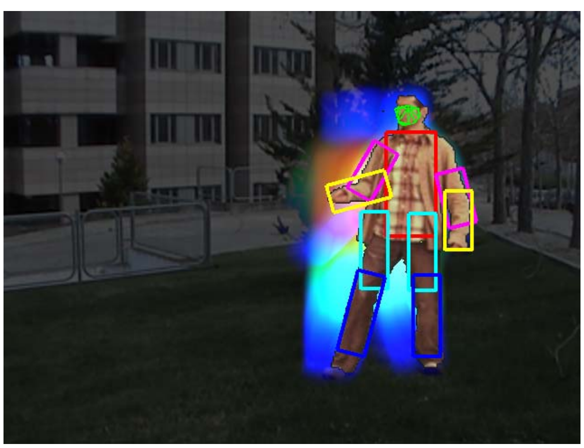


Figure 8. Application of the whole framework (pose and face recovery)

References

[1] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary region segmentation of objects in n-d images, 2001. 2

[2] T. Cootes, J. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685. 4

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59. 4

[4] D. Corrigan, S. Robinson, and A. Kokaram. Video matting using motion extended grabcut. *IET Conference Publications*, 2008. 1

[5] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. volume 2, pages 886–893, 2005. 1, 2

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1). 4

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1, 4, 5

[8] D. Geronimo, A. Lopez, and A. Sappa. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 1

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. (07-49), October 2007. 4

[10] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 1, 2, 4, 5, 7

[11] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. 23(3):309–314, 2004. 1, 2

[12] F. Tiburzi, M. Escudero, J. Bescos, and J. Martinez. A ground-truth for motion-based video-object segmentation: http://www-gti.ii.uam.es/cvsg. *IEEE International Conference on Image Processing (Workshop on Multimedia Information Retrieval*, 2008. 5

[13] P. Viola and M. J. Jones. Robust real-time face detection. volume 57, pages 137–154, 2004. 1, 3