Semi-Automatic Object Tracking in Video Sequences

Federico Lecumberry
I.I.E. - Fac. de Ingeniería - Univ. de la República
Montevideo, CC30, Uruguay
and
Alvaro Pardo
I.I.E. - Fac. de Ingeniería - Univ. de la República
D.I.E. - Univ. Católica del Uruguay
Montevideo, CC30, Uruguay*

Abstract

A method is presented for semi-automatic object tracking in video sequences using multiple features and a method for probabilistic relaxation to improve the tracking results producing smooth and accurate tracked borders. Starting from a given initial position of the object in the first frame the proposed method automatically tracks the object in the sequence modelling the a posteriori probabilities of a set of features such as color, position and motion, depth, etc.

Keywords: signal processing, video, segmentation, objects, recognition.

1 Introduction

In the past, several authors have proposed to solve the problem of object tracking and segmentation using color, texture or motion information alone [9]. It can be shown that no single visual feature can be enough to successfully solve the problem in the wide variety of real world scenes. For example, the color of the object can be similar to part of the background and spatial information is needed to overcome this ambiguity. In addition, it is well known that not all features perform uniformly in every situation: motion and texture features are computed using neighborhood operations that tend to be unreliable at object boundaries. On the other hand, motion alone can discriminate between regions of similar color undergoing different motion.

At the light of these observations, the combination of features emerged as a promising framework. For instance, are referred the combination of color and spatial information [3, 5, 13], color, spatial and motion information [8, 2], color and depth [6], etc. The use of multiple features not only gives more information but more specifically, it provides with some complementary information. Khan and Shah [8] proposed to join optical flow, color and spatial information into a single feature vector and then build a Gaussian Mixture Model (GMM). To classify the pixels they compute a weighted sum of individual feature likelihoods. The problem with this approach is that motion is very noisy and plays a minor role in the final segmentation while adding noisy estimations. In the present work position, color and motion information is separated. Parametric models are applied to motion estimation, which are then used to update the object position in the image.

Traditionally, probabilistic modelling has been used in order to transform the segmentation problem into a classification one. The main goal is to model the PDF of each feature and then apply the maximum a posteriori (MAP) principle to classify individual pixels to the corresponding class. Let assume that there are two possible classes from the set $\{O, B\}$ where O stands for the object of interest and B for the background, and a set of features $\{f_1, ..., f_N\}$. Then, if independence is assumed, the a posteriori probabilities $P(\chi|f_1, ..., f_N)$ with $\chi \in \{O, B\}$ can be written as:

$$\prod_{i=1}^{N} P(\chi|f_i) \tag{1}$$

That means that independent information is combined via the product of individual a posteriori probabilities for each feature. In other works [8, 11, 7], instead of applying Eq. (1) the combination is done using a weighted sum of the a posteriori probabilities $P(\chi|f_i)$:

$$\sum_{i=1}^{N} \omega_i P(\chi|f_i) \tag{2}$$

This also allows us to introduce the confidence of each measure in the weighting factors ω_i. In [12] the authors studied the problem of classifier combination by averaging and multiplying. They concluded that the averaging classifier is to be preferred in the case when posterior probabilities contain errors. On the other hand, the product rule outperforms averaging when the poste-

*This work was funded by a grant PDT S/C/OP/17/07 and the "IIE-Tecnocom" posgraduate scholarship.

rior probabilities are accurately estimated. The underlying assumption is that averaging reduces the estimation errors. In addition, the statistical dependence of the features must be considered. If the features are independent and computed without errors, the product rule should be used to take advantage of the independent representations. In the case of noisy estimations, the average rule should be preferred. Given that the estimations from different features contain errors, such as mismatch between the observed color and the modelled ones, errors in the motion estimation, etc., the average rule is used in this work.

The outline of the present paper is as follows: section 2 presents a method for probability relaxation that will be used to smooth the posterior probabilities, section 3 describes the method and the features used in the tests. Section 4 presents the results and finally section 5 presents the conclusions and future work.

2 Modified Vector Probability Diffusion

In [10] a method for the diffusion of probability vectors was introduced. Given a vector of probabilities $p(x) \in \mathcal{P} = \{p \in \mathbb{R}^m : \|p\|_1 = 1, p_i \geq 0\}$, the anisotropic diffusion that minimizes the L_1 norm, $\int \|\nabla p\|$, of this vector restricted to the probability simplex $\mathcal{P}$ is:

$$\frac{\partial p_i}{\partial t} = \nabla . \left(\frac{\nabla p_i}{\|\nabla p\|} \right) \quad i = 1, ..., m \tag{3}$$

This evolution equation slows the diffusion at points with high $||\nabla p|| = \sqrt{\sum_{i=1}^{m} ||\nabla p_i||^2}$. A modification of Eq. (3) is presented in order to stop the diffusion along a desired direction z while allowing the diffusion in the normal direction. The main goal is to inhibit the diffusion across the object borders while allowing the diffusion along them.

Let z be a vector field with unit length and direction normal the object borders. For each component p_i of the probability vector, the direction of diffusion in Eq. (3) is ∇p_i[1]. Since the diffusion is intended to stop across the object border, the direction of diffusion is modified subtracting the component across the border, i.e. the component parallel to z. Hence, the new diffusion direction is defined as: $\nabla p_i - \langle z, \nabla p_i \rangle z$. To correct the strength of the diffusion the norm of the gradient must be modified to suppress the contribution of the derivative parallel to z. Taking into account the previous modifications, the corresponding Modified Vector Probability Diffusion (MVPD) equation becomes:

$$\frac{\partial p_i}{\partial t} = \nabla . \left(\frac{\nabla p_i - \langle \nabla p_i, z \rangle z}{\|\nabla p - \langle z, \nabla p \rangle z\|} \right) \tag{4}$$

for $i = 1, ..., m$. To select the object borders z is set as $z = \left(\frac{u_x}{\sqrt{b^2 + u_x^2 + u_y^2}}, \frac{u_y}{\sqrt{b^2 + u_x^2 + u_y^2}} \right)$, where u is the luminance component of the image and b is a parameter that selects the relevant borders as points with $||\nabla u(x, y)|| \gg b$.

It can be easily shown that the Eq. (4) comes from the minimization of the functional:

$$\int \|\nabla p_i - \langle z, \nabla p_i \rangle z\| \, dx$$

Furthermore, it can be shown that the evolution guarantees that the probability lives in the manifold of vectors with components adding up to one and its discrete version fulfills a maximum principle and in this way the diffusion remains in the probability simplex.

The numerical implementation of Eq. (4) can be done using standard numerical methods taking forward differences for the gradients and backward differences for the divergence operator. The stopping time is automatically selected when the L_1 norm of the result is a fraction of the L_1 norm of the initial condition.

3 The method

This section presents the features used in this work: color, position (updated via motion estimation), and depth, and the posterior probability estimation and combination.

Color Color is an interesting feature for deformable object tracking since it is robust against different types of deformations. Color is represented in the L*a*b* color space and model object and background color pdfs with a GMM:

$$p(f_c|\chi) = \sum_{i=1}^{n_c} \alpha_i^c \mathcal{N}_i(\mu_i^c, \Sigma_i^c)$$

where $\mathcal{N}(\mu_i, \Sigma_i)$ is a gaussian kernel with mean μ_i and covariance Σ_i.

Despite its robustness, color information alone is generally not enough to completely resolve the spatial localization and shape of the object. Some authors proposed to include the (x, y) pixel positions in the feature vector [3, 5]. Including the pixel position into the feature vector produces compact clusters. However, when using GMM these clusters are elliptical and restrict the type of objects that can be modeled. Therefore, as in [13], the spatial/position distribution is computed separately using kernel methods. Motion estimation is used to obtain an estimation of the object shape.

[1] This comes from the fact that the heat equation $\frac{\partial f}{\partial t} = \nabla.(k \nabla f)$ diffuses the heat f in the direction of ∇f with a conduction coefficient k.

Motion Motion information is taken into account to estimate the object shape at frame $t+1$, $\hat{S}(t+1)$, given the shape at frame t, $S(t)$. To track the object shape deformation an optical flow technique is applied to obtain the affine motion of the object [14]. The optical flow $(v_x(x,y), v_y(x,y))$ is obtained as the linear least square solution of:

$$\sum_{(x,y)\in S(t)} [u(x-v_x, y-v_y; t+\Delta t) - u(x,y;t)]^2 \tag{5}$$

Taking a first order approximation of Eq. (5) and using:

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} a_1 + a_2 x + a_3 y \\ b_1 + b_2 x + b_3 y \end{bmatrix}$$

a system of linear equations is obtained for the parameters $\{a_1, a_2, a_3, b_1, b_2, b_3\}$. To add robustness to the estimations, in the previous summations only the points in $S(t)$ where the gradient can be computed with confidence are considered. In the implementation are only considered points (x,y) such that $||\nabla u(x,y)|| > 16$.

Position The object position probability, $P(O|f_s)$, is computed convolving the position support of the object, $S(t)$, with a Gaussian kernel, and the background probability is then $P(B|f_s) = 1 - P(O|f_s)$.

Depth In the case of the flowers sequence (in figure 6) also is included the depth feature via the disparity. The disparity for the i-th frame is calculated using the algorithm proposed by Bobick and Intille [1]; considering the i-th and $(i+1)$-th frames as the left and right images of a stereo rig. Even though this sequence was not prepared for stereo processing the apparent object motion (front-parallel to the camera plane)[2] and the relative depth difference of the objects of interest, allows the use of this algorithm to estimate the disparity; which is verified with the experimental disparity estimation (see figure 6).

To estimate a probability function for each pixel given its disparity value, a histogram model is used for the background and object updating this model in each frame.

Posterior probabilities combination With the a posteriori probabilities for each feature the total posterior probabilities are estimated as[3]:

$$\hat{P}(\chi) = \sum_{i=1}^{N} \omega_i P(\chi|f_i) \tag{6}$$

[2] Which guarantee that same rows of consecutive frames are "closely" in epipolar correspondence.

[3] In the implementation weights are selected uniform $\omega_i = 1/N$.

Finally, contextual information is introduced applying MVPD to the vector of posterior probabilities $(\hat{P}(O), \hat{P}(B))$, before pixel-wise MAP classification. This helps to overcome errors during the posterior probability estimation and adds coherence to the results.

Algorithm The only input of the algorithm is the initial object position, $S(0)$. From it, the posterior probabilities for each feature are computed. In the case of color, the initial condition of the EM algorithm is obtained with the fuzzy C-means algorithm. The optimal number of components in the GMM is automatically selected using the modified EM method proposed in [4]. Then, given $S(t)$, the algorithm proceed as follows:

1. Estimate the new object shape, $\hat{S}(t+1)$, using motion information.
2. Compute the posterior probabilities for the selected features (shape, color, disparity, etc.).
3. Obtain the posteriori probabilities (Eq. (6)).
4. Diffuse the posteriori probabilities (Eq. (4)).
5. Obtain the new object position using MAP rule.

4 Results

In the first experiment the results of the proposed algorithm using MVPD against the same algorithm using VPD and the algorithm without probabilistic relaxation (WOP) are compared. To assess the quality of the results the number of false positives (FP) and false negatives (FN) (figure 2(a) and 2(b) respectively) are computed with respect to a sequence segmented by hand by an experienced user.

Figures 2(c) and 2(d) summarize the results from two representative frames. It is remarkable how the results with diffusion reduce the number of FN due to the regularization at the borders and a small increase of the FP. In addition, it can be observed that the results with MVPD reduce even more the number of FN. That means that with respect to the results expected by the expert the algorithm improves via the use of probability diffusion, particularly MVPD.

Figure 2(c) and 2(d) show the localization of FP (in black) and FN (in white) for MVPD and VPD. To understand the increase of FP the following facts must be considered. First, the user consistently omitted the hair close to the neck while the methods with diffusion (MVPD and VPD) included it; these points constitute the FP close to

the neck. This explains the almost constant differences between the results with VPD or MVPD, and WOP. Second, the hand-segmented results have rugged borders and the results with diffusion smooth ones.

The FN, mainly concentrated at the helmet, are points with colors similar to the ones of the background, and also weak borders (gradients). That means that the hand segmented image has a global subjective decision that is not taken into account explicitly in the algorithm. Even though these characteristics of the video sequence, the algorithm successfully tracks the helmet border across several frames.

Figure 3 shows that VPD and MVPD improve the smoothness of the object borders whit respect to WOP. MVPD smoothes even more while respecting the borders. Since these improvements can only be seen at the borders, the improvements are small in terms of FN. To remark the differences between VPD and MVPD figure 1 show the diffused $\hat{P}(O)$ for each method.

To conclude, qualitatively can be observed that the results after MVPD are smooth and the number of FN decreases at the cost of a slightly increasing of the FP while obtaining a stable segmentation across several frames, up to frame 200.

In the second example in figures 4 and 5, the results for the segmentation of sequences `foreman` and `carphone` are presented. The results are stable and precise across several frames up to frame 300. In `carphone` the method successfully tracks the object capturing the motion, deformation and zooming. In frame 178 another object (the hand) with similar features occludes the main object (head) and for this reason the algorithm included it as part of the object being segmented. These kind of problems were not modelled in the proposed method. Hence, the results were as expected. Later, when the hand disappears form the scene the method segments only the head.

In the third example in figure 6, the results of the disparity estimation and the segmentation results with the combination of color, position and depth (disparity) are presented. The second row presents the results using the method described above. In the results for the frames 6 and 21 part of the background is included as part of the object. This is due to the closeness of the color and position features of background and object. For the same reason in frame 38 the object includes the branches. In this case the result is correct. The third column presents the results using only depth and color features[4]. Here the depth corrects the problems commented before but the algorithm used for the depth estimation introduces other errors that deteriorate the segmentation, see at the left of the tree at frames 21 and 38. Finally, in the last row, all the previously problems are solved integrating in the proposed method color, position and depth features. Although some problems remain present, the improvements achieved due the combination of several features are presented.

Finally, the fourth example in figures 2(e)-2(h) shows an example changing the weights of color and position features. In figures 2(e) and 2(f), the color and position have the same weight and the chimney gets lost during the first frames because the color model cannot discriminate between object and background. Figures 2(g) and 2(h) shows how this can be solved using different weights assigning more weight to the position then the color, $\omega_p > \omega_c$.

[4]Remember that always is used the motion feature for the update of position.

5 Conclusions

This work presents an algorithm for semi-automatic object tracking in video sequences using several features and a new probabilistic relaxation method (a modified version of an existing VPD). Although the results presented can be improved with more sophisticated methods, it is showed that a simple method together with MVPD is able to track objects in long sequences producing smooth and accurate borders.

Even though the proposed method is a region-based one and no constraints are imposed to the boundary of the tracked regions, the borders are smooth. The accuracy of the borders of the tracked objects depend on the power of discrimination of the selected features, and the appearance of new objects and/or background. The algorithm does not consider the latter. In order to overcome some of these limitations is planed to use snakes or other methods to further improve the results.

The tracked objects in the examples are almost rigid objects. For the case of non-rigid objects the motion estimation must be improved to track the deformations of the object. The solution might be to apply the same motion estimation on a region basis. For example, dividing the object in small regions and then applying a grouping principle.

The results and the full color images can be obtained from this web site: http://iie.fing.edu.uy/investigacion/grupos/gmm/.

References

[1] A. F. Bobick and S. S. Intille. Large Occlusion Stereo. *Int. Journal of Computer Vi-*

sion, 33(3):181–200, September 1999.

[2] R. Castagno, T. Ebrahimi, and M. Kunt. Video Segmentation Based on Multiple Features for Interactive Multimedia Applications. *IEEE Trans. on Circuits and Systems for Video Tech.*, 8(5):562–571, September 1998.

[3] M. Everingham and B. Thomas. Supervised Segmentation adn Tracking of Nonrigid Objects using a Mixture of Histograms Model. In *ICIP01 - Int. Conf. on Image Processing*, pages 62–65, 2001.

[4] M. Figueiredo and A. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Trans. on Pattern and Machine Intelligence*, 24(3):381–396, March 2002.

[5] H. Greenspan, J. Goldberger, and A. Meyer. Probabilistic Space-Time Video Modeling via Piecewise GMM. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 26(3):384–396, March 2004.

[6] M. Harville, G. Gordon, and J. Woodfill. Foreground segmentation using adaptive mixture models in color and depth. In *IEEE Workshop on detection and recognition of events in video*, pages 3–11, 2001.

[7] E. Hayman and J. Eklundh. Probabilistic and Voting Approaches to Cue Integration for Figure-Ground Segmentation. In *ECCV 2002*, LNCS 2352, pages 469–486.

[8] S. Khan and M. Shah. Object Based Segmentation of Video using Color, Motion and Saptial Information. In *CVPR2001 - Int. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 746–751, 2001.

[9] N. Friedman and S. Rusell. Image segmentation in video sequence: A probabilistic approach. In *Conf. Uncertainty in Artificial Intelligence*, number 13, 1997.

[10] A. Pardo and G. Sapiro. Vector Probability Diffusion. *IEEE Signal Processing Letters*, 8(4):106–109, April 2001.

[11] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Machine Vision and Applications*, 14:50–58, 2003.

[12] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33:1475–1485, 2000.

[13] D. Thirde, G. Jones, and J. Flack. Spatio-Temporal Semantic Object Segmentation using Probabilistic Sub-Object Regions. In *BMVC2003*, 2003.

[14] J. Wang and E. Adelson. Representing moving images with layers. *IEEE Trans. on Image Processing*, 3(5):625–638, September 1994.

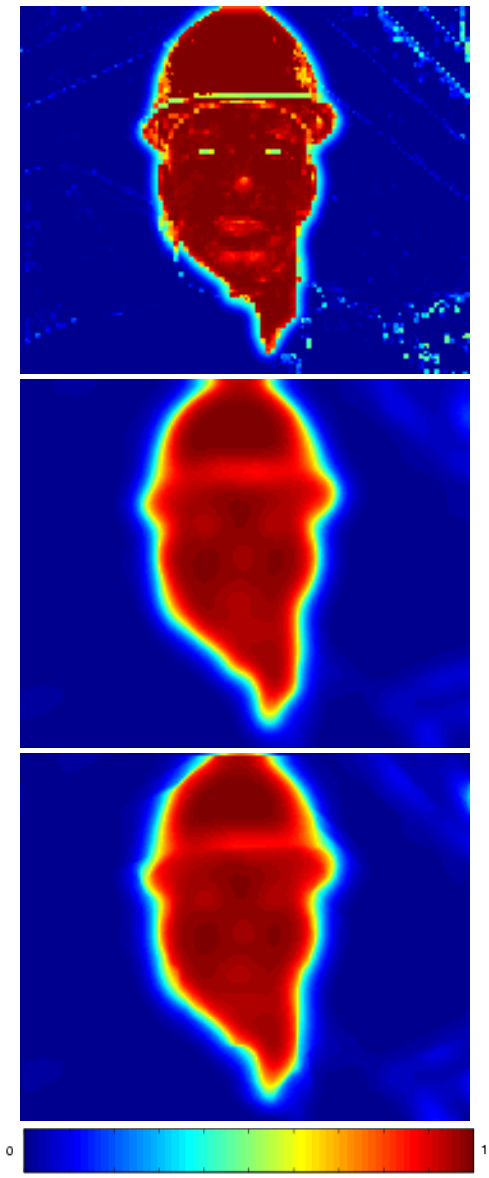
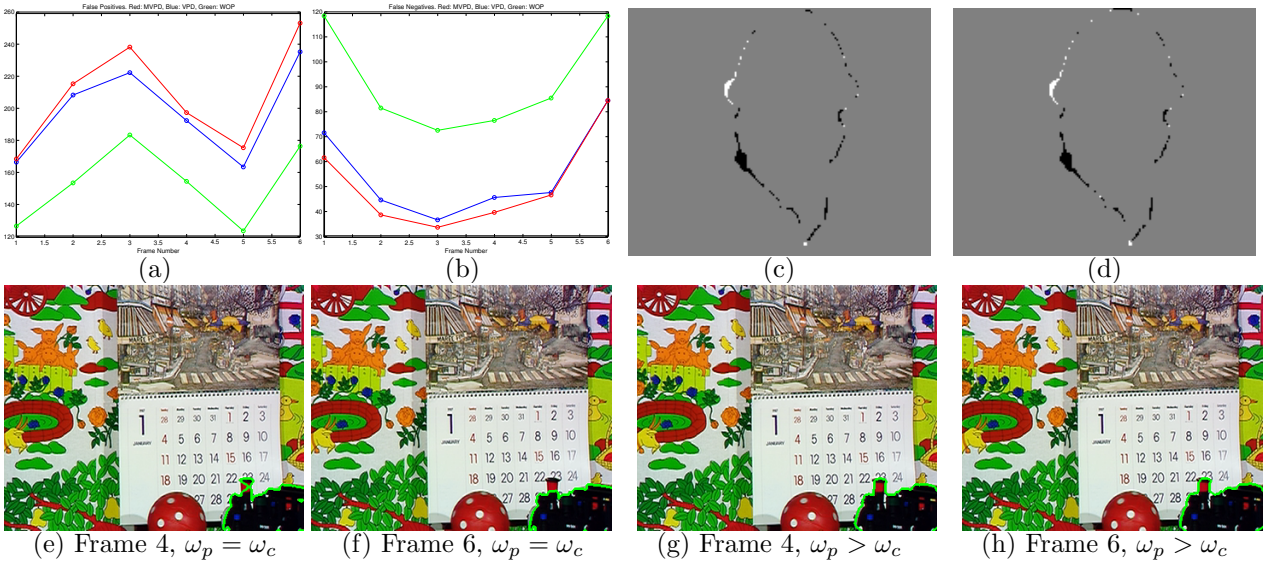


Figure 1: Results for the object probability $\hat{P}(O)$ (from top to bottom) without diffusion with VPD and with MVPD.



(a) (b) (c) (d)

(e) Frame 4, $\omega_p = \omega_c$ (f) Frame 6, $\omega_p = \omega_c$ (g) Frame 4, $\omega_p > \omega_c$ (h) Frame 6, $\omega_p > \omega_c$

Figure 2: (a)-(b) FP and FN for frames 5, 10, 15, 50, 100 and 200. (c)-(d) FP (in black) and FN (in white) for (c) MVPD and (d) VPD. (e)-(h) Example with different weights in Eq. 2.

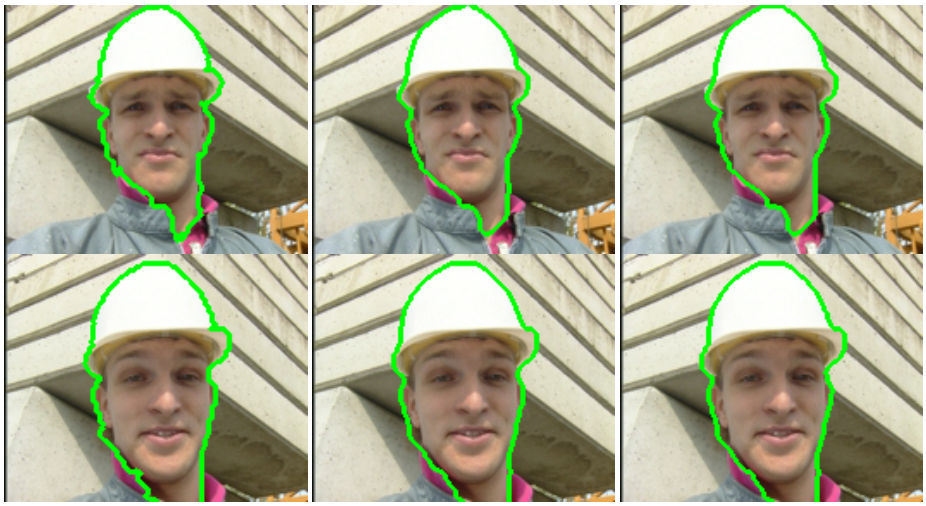


Figure 3: Left: Results of WOP. Middle: Results with VPD. Right: Results with MVPD. Up: frame 5. Bottom: frame 100.

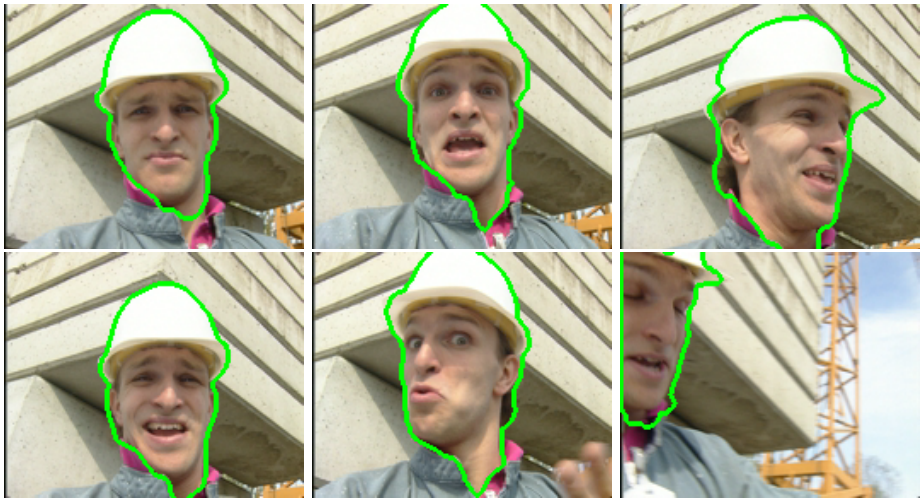


Figure 4: Results for `foreman` sequence with color, position and MVPD. Frames 2, 66, 111, 122, 191 and 286.

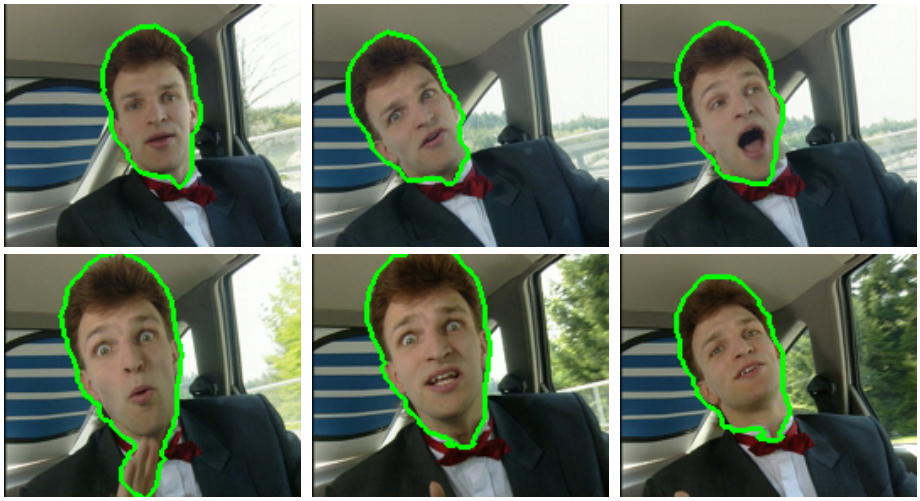


Figure 5: Results for `carphone` sequence with color, position and MVPD. Frames 2, 89, 115, 178, 187 and 300.

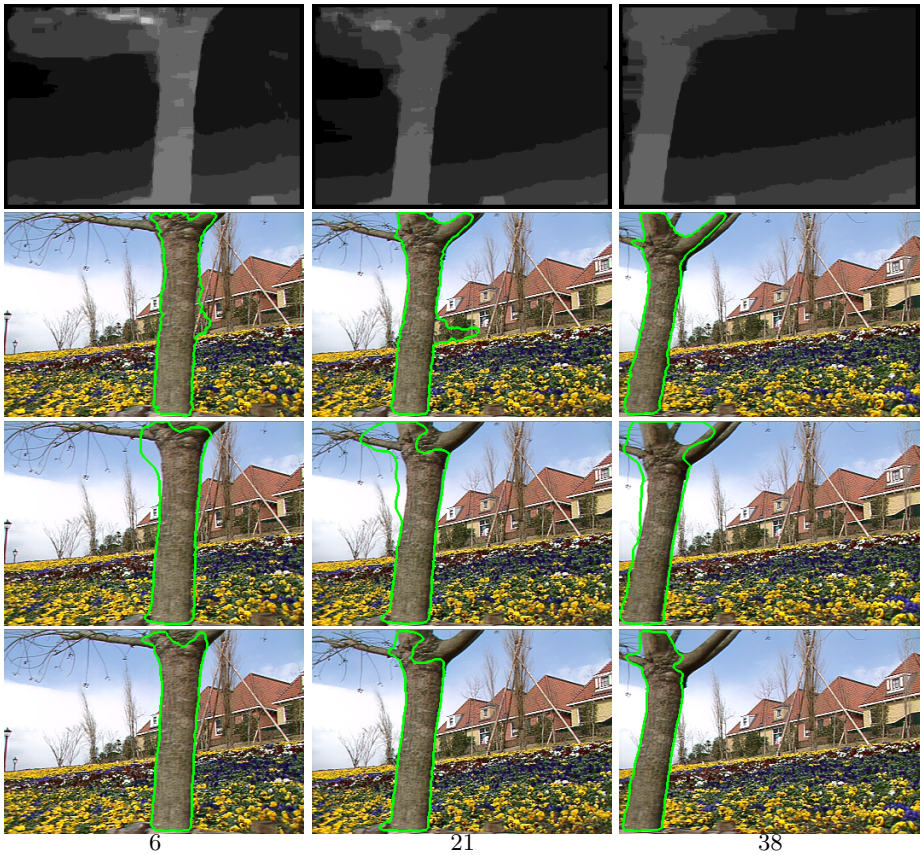


Figure 6: From top to bottom: Disparity estimation. Results using color and position with MVPD. Results using position and depth with MVPD. Results using position, depth and color with MVPD. Frames 6, 21 and 38.