BMVC 2002

Multiple Plane Segmentation Using Optical Flow

Marco Zucchelli† José Santos-Victor‡ Henrik I. Christensen†
† CVAP & CAS, Royal Institute of Technology, Stockholm, Sweden S100 44
‡ Vislab, Instituto Superior Técnico, Lisboa, Portugal

Abstract

In this paper we present a motion based segmentation algorithm to automatically detect multiple planes from sparse optical ¤ow information. An optimal estimate for planar motion in the presence of additive Gaussian noise is £rst proposed, including directional uncertainty of the measurements (thus coping with the aperture problem) and a multi-frame ($n > 2$) setting (adding overall robustness). In the presence of multiple planes in motion, the residuals of the motion estimation model are used in a clustering algorithm to segment the different planes. The image motion parameters are used to £nd an initial cluster of features be-longing to a surface, which is then grown towards the surface borders. Initialization is random and only robust statistics and continuity constraints are used. There is no need for using and tuning thresholds. Since the exact parametric planar ¤ow model is used, the algorithm is able to cope ef£ciently with projective distortions and 3D motion and structure can be directly estimated.

1 Introduction

Planes are common features in both man made and natural environments. The underlaying geometry of induced homographies is well understood and used to perform different tasks: video stabilization, visualization, 3D analysis (using for example plane + parallax [1]), egomotion estimation, calibration, just to cite few of them. In the continuous limit the homography is replaced by the Flow Matrix [3]: this can be calculated from two views [4], or as suggested more recently from Irani [5] in a multi-frame context to gain more stability and precision.

Automatic detection of planar surfaces from ¤ow £elds belongs to the wider area of motion based segmentation, where the image is partitioned into regions of homogeneous 2D motion based on continuity or on £tting a parametric motion model. There exist quite a large amount of different approaches to the solution of this problem. *Top-down* techniques handle the whole image as the estimation support. A global motion model is estimated and areas that do not conform to such a model are detected, generating a two class partition [6]. The main limitations are (*i*) the presence of a dominant motion is required and (*ii*) simply rejecting non conforming pixels does not produce spatially compact regions. Segmentation in a Markovian framework enables addition of spatial consistent constraints [7]. Simultaneous estimation of models and supports is another approach useful when a *mixture of motion models*, none of them dominant, is present. The EM algorithm has been used ef£ciently for this purpose in [8]. A more general approach consists in partitioning the image into *elementary regions* (intensity-based, texture-based regions or square blocks are often used) and searching motion based regions as clusters of these. A commonly used technique consists in £tting af£ne motion to the regions and grouping them with a clustering process based on similarity of the model parameters. In [9] a *k-means* algorithm was used in the motion parameters space.

BMVC 2002 doi:10.5244/C.16.29

In this paper we present an automatic clustering technique that works on a sparse ¤ow £eld and is able to £nd features laying on planar surfaces by analyzing the ¤ow matrix that they generate. There are three main contributions:

- We £rst show that greater robustness in the estimation of the planar ¤ow parameters can be achieved by re-weighting the linear least squares estimation by optical ¤ow covariance matrix. Due to the aperture problem, errors in optical ¤ow computation are rarely symmetric, but tend instead to be anisotropic and correlated along x and y directions. In this case re-weighted least squares is the maximum likelihood estimator. We show how to compute the covariance matrix directly from image gray levels.

- Further robustness can be obtained in the case in which multiple frames are available. In such case, the fact that the underlying planar geometry is the same in all the views, provides a rank constraint over the matrix obtained by stacking together the planar ¤ow parameters for the couples of frames.

- Planar ¤ow is £tted to a cluster of points and the standard deviation of the residuals is estimated by means of robust statistic. The consistency of each single feature image motion with the planar hypothesis can now be established by comparing its residual with the estimated standard deviation; features that have residuals larger than $2.5\hat{\sigma}$ are discarded as outliers. Clusters of inliers so obtained are grown outwards as the model allows using a nearest neighbor algorithm that takes into account continuity (i.e. points closer to a cluster are more likely to belong to it than others) and robustness (i.e. is better to begin to grow the algorithm in regions where ¤ow is estimated robustly). Since the exact parametric planar model is used, 3D motion and structure can be estimated directly from the segmentation information. As only robust statistic is used, there is no need of tuning thresholds, eliminating the need for user interaction.

Unlike top-down techniques no dominant motion is required, *unlike* af£ne ¤ow £tting, the method we propose is able to cope ef£ciently with projective distortions and works well when objects are close to the camera.

Although random sampling and robust statistics are used, *this is not a RANSAC-like algorithm* for the reason that we do not assume to have a dominant motion model. In general, we allow any relative dimensions of planes, while RANSAC is based on a high inliers to outliers ratio [10].

2 Planar Motion Estimation

2.1 Basic Model and Notation

The image motion of points on a planar surface, between two image frames can be expressed as [3]:

$$\mathbf{u}(\mathbf{x}) = F(\mathbf{x}) \cdot \mathbf{b} + \mathbf{n}(\mathbf{x}) \tag{1}$$

where $F(\mathbf{x})$ is a 2×8 matrix depending only on the pixel coordinates $\mathbf{x} = (x, y)$:

$$F(\mathbf{x}) = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{pmatrix} \tag{2}$$

$\mathbf{n} \sim N(0, \sigma)$ is Gaussian additive noise and $\mathbf{b}_{8\times 1}$ is the vector of the planar ¤ow parameters vector. The vector $\mathbf{b}$ can be factorized into a *shape* and a *motion* part as:

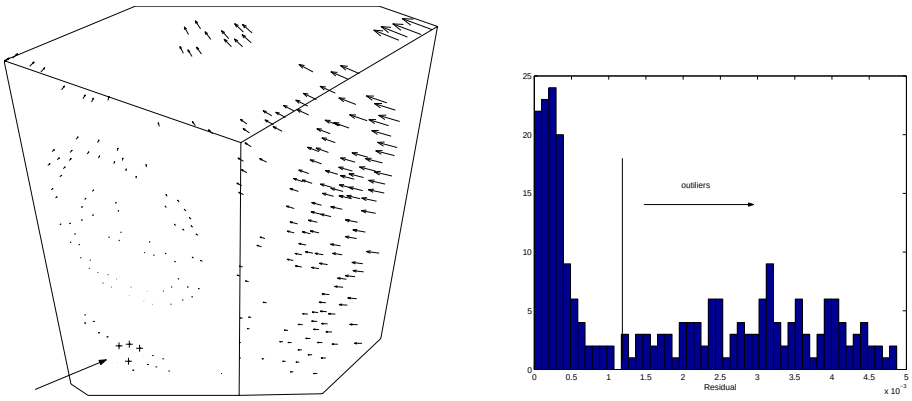
$$\mathbf{b} = S_{8\times 6} \cdot \begin{pmatrix} \mathbf{v} \\ \omega \end{pmatrix} \tag{3}$$

with

$$S = \begin{pmatrix} fp_z & 0 & 0 & 0 & f & 0 \\ p_x & 0 & -p_z & 0 & 0 & 0 \\ p_y & 0 & 0 & 0 & 0 & -1 \\ 0 & fp_z & 0 & -f & 0 & 0 \\ 0 & p_x & 0 & 0 & 0 & 1 \\ 0 & p_y & -p_z & 0 & 0 & 0 \\ 0 & 0 & -\frac{p_x}{f} & 0 & \frac{1}{f} & 0 \\ 0 & 0 & -\frac{p_y}{f} & -\frac{1}{f} & 0 & 0 \end{pmatrix} \tag{4}$$

where $\mathbf{p} = (p_x, p_y, p_z)$ is the normal to the plane, $\omega = (\omega_x, \omega_y, \omega_z)$ and $\mathbf{v} = (v_x, v_y, v_z)$ are the rotational and the linear velocities of the camera. The camera focal length is denoted by f and we assume that it is constant over time.

2.2 Two Frames Re-weighted Estimation



a b

Figure 1: (a) Optical ¤ow warping respect to 4 points on the plane chosen according to the nearest neighbor principle (the 4 points are marked with crosses). (b) Magnitude of the residuals.

If N features are available, stacking the optical ¤ow vectors $\mathbf{u}_i$ in the vector $\mathbf{U}_{2N\times 1}$, the $F(\mathbf{x}_i)$ into the matrix $G_{2N\times 8}$ and the noise $\mathbf{n}$ into the vector $\eta_{2n\times 1}$, Eq. (1) can be rewritten as:

$$\mathbf{U} = G\mathbf{b} + \eta \tag{5}$$

The maximum likelihood estimation of the planar ¤ow parameters vector $\mathbf{b}$ is given by the *weighted* linear least squares problem:

$$\hat{\mathbf{b}} = arg \min_{\mathbf{b}} \{(\mathbf{U} - G\mathbf{b})^T W(\mathbf{U} - G\mathbf{b}))\} \tag{6}$$

The solution of the LLSE problem (6) is found by solving the re-weighted system of normal equations:

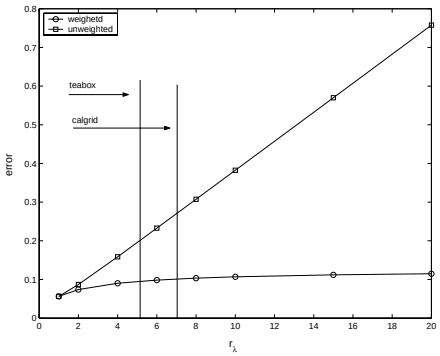


Figure 2: Performance of weighted linear least squares for the estimation of the ¤ow matrix. The average values of r_λ for two of the sequences used for tests are also reported.

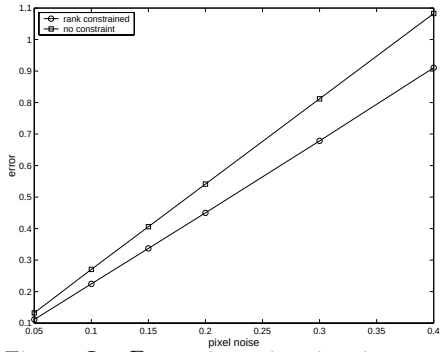


Figure 3: B matrix estimation improvement due to multi-frames integration. 10 frames were used.

$$G^T W G \hat{\mathbf{b}} = G^T W \mathbf{U} \tag{7}$$

The weight matrix W is block diagonal and the diagonal blocks are 2×2 matrices that represent the covariance of the estimated optical ¤ow vectors. The covariance matrix Σ can be estimated (see [11]) as:

$$\Sigma^{-1} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix} \tag{8}$$

The introduction of re-weighting in Equation (6) is particularly important since, due to the aperture problem, the errors are usually asymmetric. This approach has £rst been used successfully by [12] in the context of orthographic factorization.

A set of simulated tests was performed in order to assess the improvements of this approach in the estimation of the planar ¤ow parameters. We randomly generated points on a planar surface and then added Gaussian elliptical noise of randomly generated directions to the ¤ow vectors. The shape of the elliptical uncertainty was varied changing the value of the parameter $r_\lambda = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ where λ_{max} and λ_{min} are the largest and smallest eigenvalues of the covariance matrix Σ. We ran 20 trials for each of the values of r_λ for 100 points on plane. We de£ned the residual £eld as:

$$\mathbf{r} = \mathbf{u} - F\hat{\mathbf{b}} \tag{9}$$

The estimation error we used is then:

$$err = \underset{i}{mean}\left(\frac{\|\mathbf{r}_i\|_\Sigma}{\|\mathbf{u}_i\|}\right) \tag{10}$$

where $i = 1 \ldots N$ runs over the set features and $\|\|_\Sigma$ is the Σ-norm. Results reported in Figure 2 show clearly the superiority of the weighted approach.

2.3 Multi-frames Re-weighted Estimation

If m views of the same planar surface are acquired, $m-1$ pairs can be formed between the £rst and the j_{th} images, $j \in \{2, \ldots, m\}$. If such pairs are close enough such that the instantaneous approximation can be used, the set of ¤ow parameters vectors $\mathbf{b}_j$ can still be estimated by arranging them in a matrix $B_{8\times m-1} = [\mathbf{b}_2, ..., \mathbf{b}_m]$ and solving:

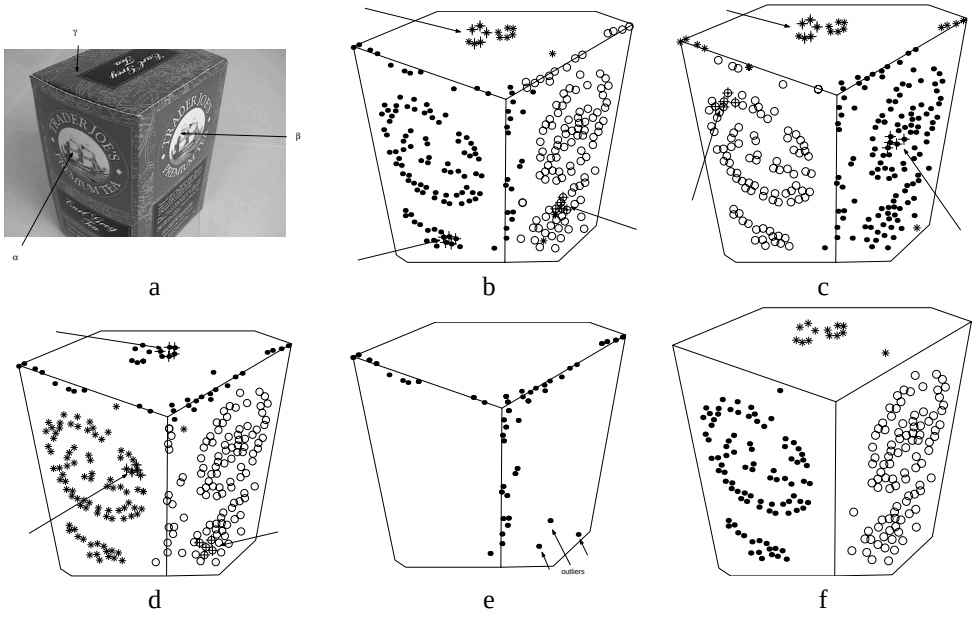


Figure 4: The *tea box* sequence. (**a**) An Image from the sequence. (**b**) First Iteration of the clustering algorithm. Crosses indicate the initial 5 points selected. (**c**) Segmentation obtained starting from surface number 2. (**d**) Segmentation obtained starting from surface number 3. (**e**) Ambiguous features. (**f**) Segmentation after ambiguous features removal.

$$G^T W G \cdot B = G^T W [\mathbf{U}_2, ..., \mathbf{U}_m] \tag{11}$$

or in short $C \cdot B = K$. The matrix G depends only on the features positions in the £rst frame and is de£ned in the previous section. The weight matrix W is estimated doing a temporal average over all the m frames. In this multi-frame setting B can be factorized as:

$$B = S_{8\times 6} \cdot \begin{pmatrix} \mathbf{v}_2 & \dots & \mathbf{v}_m \\ \omega_2 & \dots & \omega_m \end{pmatrix}_{6\times m-1} \tag{12}$$

Solving for B in Eq. (11) is equivalent to solving independently for the $\mathbf{b}_j$ of the image pairs. This does not exploit the fact that the underlying plane geometry (expressed by the matrix S) must be the same in all the views. Such constraint is expressed by the dimensionality of the matrices on the right side of Eq. (12) that £xes the rank of B to be smaller than 6. Lower ranks can be generated by special motion con£guration, for example $rank(B) = 1$ when the motion is constant over time.

Due to the fact that $rank(B) \leq 6$ we get that $rank(K) \leq 6$. Hence before solving Eq. (11) we can re-project K over a lower dimensional linear subspace seeking the matrix $\tilde{K}$ with $rank(\tilde{K}) \leq 6$ that is closest to K in Frobenius norm [3].

Figure 3 shows the error in the estimation of B for a simulated data set. A set of 10 views of the same plane was generated and matrix B estimated using Eq. (11) with or without using the rank constraint. A total of 100 features and 20 trials were used. The estimation error was de£ned as $err = \frac{\|B-\tilde{B}\|}{\|B\|}$ where B is the ground truth. Error over optical ¤ow was varied between 0.05 and 0.4 pixels. The rank constraint clearly increases the performance of the estimation. The *un-weighted* multi-frame approach was £rst used in [5].

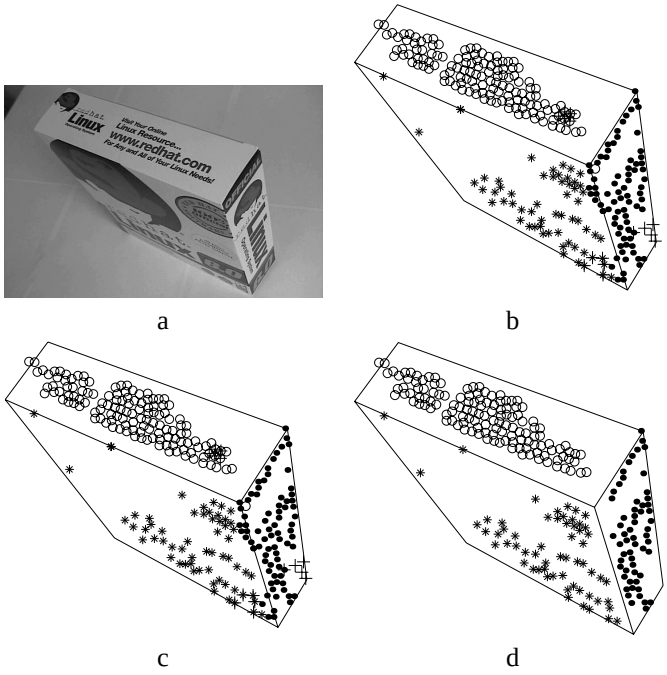


Figure 5: The *Linux box* sequence. (**a**) A frame from the video sequence. (**b**) Estimated optical ¤ow. (**c**) Segmentation obtained at the £rst iteration. 4 of the 5 initial features (crosses) on the surface nr.1 (dots) are rejected as too noisy. (**d**) Segmentation after ambiguous features removal.

3 Planar Motion Segmentation

We have proposed an *optimal* solution to the problem of estimation of planar motion in the presence of Gaussian additive noise. This method is now used as the core step of the segmentation algorithm by £tting planar motion to a tentative cluster of points and rejecting outliers by means of robust statistics.

3.1 Algorithm

The magnitude of the residuals $r = \|\mathbf{r}\|$ can be effectively used for segmentation purposes. Figure 1 (a) shows the residual ¤ow when the planar ¤ow parameters are estimated from the minimal con£guration of 4 points in the highlighted plane. A histogram of the norms of the residual vectors is plotted in Figure 1 (b): the difference in magnitude between points on the plane and off the plane is quite obvious. In the multi-frame setting a more robust estimate of the features residual is de£ned as:

$$\bar{r}_i = \frac{\sum_{j=1}^{m} r_{ij} e^{-d_j^2}}{\sum_{j=1}^{m} e^{-d_j^2}} \tag{13}$$

where j runs over the frames, i indexes the features and d_j is the average motion of the features between the frame j and the reference frame and measures the adequacy of the instantaneous approximation.

The selection of inliers and outliers is based on a robust standard deviation estimate [13]. If a moderate amount of outliers is present in a set of Q features, a robust estimate of the standard deviation of the residuals $\bar{r}_q$, $q \in \{1, \ldots, Q\}$ can be obtained as:

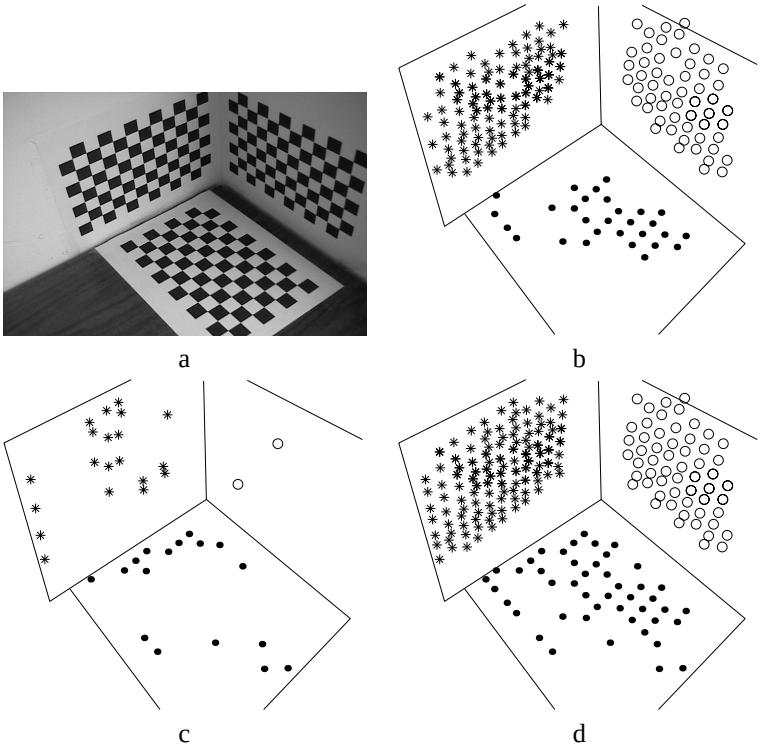


Figure 6: The *calibration grid* sequence. (a) A frame from the video sequence. (b) Segmentation obtained after three iterations of the clustering algorithm (c) Reassigned ambiguous features (d) Segmentation obtained after reassignment of the ambiguous and the rejected features.

$$\hat{\sigma} = 1.4826(1 + \frac{5}{Q-l})median\sqrt{\bar{r}_q^2} \tag{14}$$

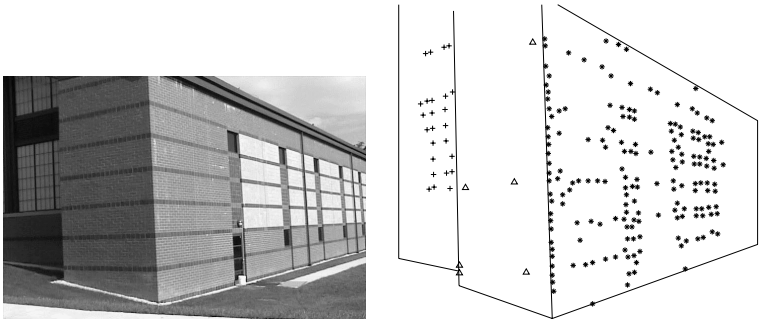
where l is the number of £tted parameters, 8 in our problem. Inliers are those that satisfy: $\bar{r}_q \leq 2.5\hat{\sigma}$.

The segmentation algorithm is outlined below.

1) **Randomly select one point** and determine an initial cluster of 5 points adding its 4 nearest neighbors. The nearest neighbor to a con£guration of points is de£ned as the point closest to the center of mass of the con£guration. The center of mass is found as a weighted mean of the features position where the weights are the minimum eigenvalue of the covariance matrix of the ¤ow vectors: this ensures that the algorithm grows, at the beginning, towards areas where the features are tracked robustly which, in turn, helps to get a more precise initial estimation of the planar ¤ow parameters. At the same time this procedure increases the probability that the initial cluster is located on a plane: it is crucial that at each step the number of outliers is small so that Eq. 14 can be used.

2) **Fit the planar ¤ow parameters** and select as good features those for which $\bar{r}_q \leq 2.5\hat{\sigma}$. If less than 4 features are, left start over; otherwise go to the next step.

3) **Add the nearest neighbor feature and start over**. Growing the cluster adding recursively the nearest neighbor exploits planes continuity, i.e. it is more likely that the



a b

Figure 7: The *aquatic center* sequence. (**a**) A frame from the video sequence. (**b**) Final segmentation. Triangles mark unclassi£ed features.

nearest neighbor to the cluster belongs to the plane than another point far away. In general, if no a priori information is available about the £lmed scene, this approach turns to be very effective.

The initial 5 points can be discarded during the growing process. This makes the algorithm more ¤exible and performing even in the case in which the initial 5 points lie on different surfaces: the *Nearest Neighbor* growing moves into one of the planar surfaces and the initial points that do not lie on such surface are later discarded as outliers. The algorithm ends naturally when all the features have been analyzed and no more inliers are found. Since the detected outliers can belong to another planar surface the algorithm can be restarted for the next plane detection.

3.2 Final Re£nement: Resolving Ambiguities

The growing algorithm can sometimes be greedy, including into a given planar area ambiguous points, which belong to a neighboring plane, but whose ¤ow is similar to that of the £rst plane. This is the case of the three sequences shown in Figures 4, 5 and 6, where planes are incident. The ambiguous points close to the intersection of the surfaces can easily be found by recursively running the clustering algorithm.

Let us call the three surfaces α, β and γ (see Figure 4 (a)) and the 3 clusters obtained running a £rst time the clustering algorithm $\mathcal{G}^1_\alpha, \mathcal{G}^1_\beta, \mathcal{G}^1_\gamma$. At this point we know approximately features belonging to the 3 planes up the ambiguous ones close to the incidence. We re-run the algorithm taking as initial feature the one closest to the center of mass of the features in $\mathcal{G}^1_\beta$. In this way the plane β is, in the new run of the algorithm, detected as the £rst one and the cluster will tend to invade the surfaces α and γ close to the borders. The algorithm £nds 3 more clusters over the 3 surfaces: $\mathcal{G}^2_\alpha, \mathcal{G}^2_\beta, \mathcal{G}^2_\gamma$. We get that the ambiguous features of surfaces α, β are de£ned as:

$$\mathcal{B}_{\alpha\beta} = \mathcal{G}^1_\alpha \cap \mathcal{G}^2_\beta \tag{15}$$

Running the algorithm a third time starting from the center of mass of the plane γ the ambiguous features close to the three intersections can be de£ned as:

$$\mathcal{B}_{\alpha\beta} = \mathcal{G}^1_\alpha \cap \mathcal{G}^2_\beta \tag{16}$$
$$\mathcal{B}_{\alpha\gamma} = \mathcal{G}^1_\alpha \cap \mathcal{G}^3_\gamma \tag{17}$$
$$\mathcal{B}_{\beta\gamma} = \mathcal{G}^2_\beta \cap \mathcal{G}^3_\gamma \tag{18}$$

The £nal clusters of points that lie on the three surfaces are de£ned as :

$$\mathcal{G}_\alpha = \mathcal{G}^1_\alpha \setminus (\mathcal{B}_{\alpha\beta} \cup \mathcal{B}_{\alpha\gamma}) \tag{19}$$
$$\mathcal{G}_\beta = \mathcal{G}^1_\beta \setminus (\mathcal{B}_{\alpha\beta} \cup \mathcal{B}_{\beta\gamma}) \tag{20}$$
$$\mathcal{G}_\gamma = \mathcal{G}^1_\gamma \setminus (\mathcal{B}_{\alpha\gamma} \cup \mathcal{B}_{\beta\gamma}) \tag{21}$$

The £nal planar ¤ow parameters $\mathbf{b}$ for the three planes are found by re£tting the clusters according to Eq. (11).

Ambiguous points can eventually be reassigned by checking their residual respect each of the £nal plane hypotheses and assigning the feature to the plane respect to which the residual is minimum and veri£es Eq. (14) where $\hat{\sigma}$ is calculated just from the points already in the cluster. Eventually, rejected points can be also assigned using this principle. The reason is that the ef£ciency of assigning a point to a cluster depends on how large the cluster is (i.e how many points the cluster has) due to the fact that the statistical precision of the £t of the B matrix grows in function of the number of points in the cluster. This means that points that are erroneously discarded when the cluster size is small can be successfully assigned when the cluster is completely growth. An example is shown in Figure 6 where we applied reassignment to the *calibration grid* sequence. Figure 7 shows an application of the algorithm to segment the surfaces of a building. Features marked with triangles are not assigned to any surface.

4 Results

We tested our algorithm extensively on real images and simulated data.

Figure 4 shows the process of iterating three times the clustering and the ambiguous features detection and removal. The process of £nding ambiguous features £nds not only features at the intersection of planes but also eventual outliers. A total of 10 frames was integrated and 237 features used; the number of features rejected was zero and 83 ambiguous features were found.

Figure 5 shows another application for a different sequence made of 11 frames and 254 features. Only 9 features were rejected and 44 features were removed as ambiguous.

Figure 6 illustrates an application to a very noisy sequence of 10 frames. Aperture problem is very serious due to the massive presence of edges (see also Figure 1). In this case we decided to reassign ambiguous and rejected features. We found that 76 of the 220 features were removed as ambiguous and all of them were reassigned correctly; 53 features were rejected of which 13 were reassigned and the reassignment was correct.

5 Conclusions

In this paper we presented a new motion based segmentation technique able to automatically £nd planar surfaces when a sparse optical ¤ow £eld is given.

We £rst formulated an *optimal* solution to the estimation of planar ¤ow parameters in presence of gaussian addictive noise. Experiments on simulated data show the improvement of performance we obtained compared with previous approaches.

We then showed how the planar geometry induces a constraint between planar ¤ow parameters estimated using different couples of frames and stated the performance improvement we can obtain by applying such constraint.

Robust estimation of planar ¤ow parameters is the core of the segmentation algorithm. A cluster of points is initialized randomly and then grown on a plane by mean of robust statistics, i.e. £nding and eliminating outliers, and proximity constraints, i.e. using the fact that planes are mostly continuous surfaces.

Results with real images were presented to illustrate the performance of the proposed method.

References

[1] Sawhney H. 3d geometry from planar parallax. Technical report, IBM Research Division.

[2] Cipolla R. Malis E. Self-calibration of zooming cameras observing an unknown planar structure. In *ICPR*, volume 1, pages 85–88, 2000.

[3] Kanatani K. *Geometric Computation for Machine Vision*. Clarendon Press, 1993.

[4] Anandan P. Bergen J. and Hanna K. Hierarchical model-based motion estimation. In *ECCV*, pages 237–252, 1992.

[5] Zelnik-Manor L. and Irani M. Multi-frame alignment of planes. In *CVPR*, volume 1, pages 151–156, 1999.

[6] Szeliski R. Torr P.H.S. and Anandan P. An integrated bayesian approach to layer extraction from image sequences. *PAMI*, 23(3):297–303, 2001.

[7] Odobez J.M. and Bouthemy P. Mrf-based motion segmentation exploiting a 2d motion model robust estimation. In *ICIP*, pages 628–631, 1995.

[8] Ayer S. and Sawhney H.S. Compact representations of videos through dominant and multiple motion estimation. *PAMI*, 18(8):814–830, 1996.

[9] Wang J. and Adelsson E. Layered representation for motion analysis. In *CVPR*, pages 361–366, 1993.

[10] Fischler M.A. and Bolles R.C. Random sample consensus: A paradigm for model £tting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, 1981.

[11] Shi J. and Tomasi C. Good features to track. In *CVPR*, pages 593–600, 1994.

[12] Irani M. and Anandan P. Factorization with uncertainty. In *ECCV*, volume 1, pages 539–553, 2000.

[13] Rousseeuw P.J. and Leroy A.M. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.