IEEE/ACM Int. Symp. on Mixed and Augmented Reality,
pp 52-55, Santa Barbara, CA, October 2006

Hybrid tracking algorithms for planar and non-planar structures subject to illumination changes

Muriel Pressigout*
Université de Rennes 1, IRISA, Lagadic project, Rennes, France

Éric Marchand†
INRIA, IRISA, Lagadic project, Rennes, France

Abstract

Augmented Reality (AR) aims to fuse a virtual world and a real one in an image stream. When considering only a vision sensor, it relies on registration techniques that have to be accurate and fast enough for on-line augmentation. This paper proposes a real-time, robust and efficient 3D model-based tracking algorithm monocular vision system. A virtual visual servoing approach is used to estimate the pose between the camera and the object. The integration of texture information in the classical non-linear edge-based pose computation provides a more reliable tracker. Several illumination models have been considered and compared to better deal with the illumination change in the scene. The method presented in this paper has been validated on several video sequences for augmented reality applications.

1 Introduction

It is important for AR applications that synthetic elements are rendered and aligned in the scene in an accurate and visually acceptable way. The registration problem is therefore a major issue. This paper addresses the problem of robust real-time model-based tracking of 3D objects using a monocular vision system: a camera. It proposes to integrate texture information in an edge-based process to get a spatio-temporal tracker that is accurate and more robust to textured environnements than classical 3D trackers. It also investigates several points about texture and the illumination problems.

In computer vision, most of the available tracking techniques can be divided into two main classes: feature-based and model-based. The former approach focuses on tracking 2D features such as geometrical primitives (points, segments, circles,...) [23], object contours [11], regions of interest [12, 8]... Such features may also be used in 3D tracking for the registration [21, 24] or help to improve tracking results as in [6, 17, 27]. A texture-based approach may suffer from lack of precision if scale changes. Another drawback of such an approach is its sensitivity to changes in illumination. We will go back on this point latter. The latter approach, the model-based one, explicitly uses a model of the tracked objects. This can be a CAD model [1, 2, 4, 14, 16] or a 2D template of the object [13]. This second class of methods usually provides a more robust solution. The most classical approaches rely on a pose mono-image estimation approach, considering that a CAD model is available. The problem is solved using registration techniques that allow alignment of 2D image data and a 3D model [18, 1, 2, 16]. Relying only on edge information provides good results when tracking sharp edges even if there are illumination changes. However, it can lead to jittering and even to erroneous pose estimation if the environment or the object is highly textured.

As one can note, model-based trackers can be mainly divided in two groups, the edge-based ones and the textured-based one, dealing with different kinds of objects or environment. However, in a realistic video sequence, the difference between each case is not so clear. Furthermore, both have complementary advantages and drawbacks. The idea is then to integrate both approaches in the same process. [19] proposes for example such a framework to estimate the image motion. [17] uses a 2D tracking based on the dominant motion estimation to initialize the 3D tracking relying on the edge projection. Merging both approaches for 3D tracking has been studied in some recent works [25, 27]. [25] fuses in a Kalman filter (EKF) framework measurements about the object's center of mass using color information, edge orientations and positions and some feature displacements obtained by a SSD minimization of the grey level difference between the current image and the prediction.

In this paper, pose and camera displacement computation is formulated in terms of a full scale non-linear optimization: Virtual Visual Servoing (VVS) [1]. In [22], the general framework fusing a model-based approach based on the edge extraction and a temporal matching relying on the texture analysis is presented. Here, we extend our previous work [1, 22] in two directions: first we consider not only piecewise planar objects (polyhedral objects) but also non-planar objects ; second since texture is considered it is also important to be robust to important illumination changes, therefore different illumination models have been studied. Many works already deal with the problems posed by changes in illumination. One can cite various works for patch tracking, from classical approaches assuming a perfect conservation of luminance to more sophisticated methods based on the photometric model presented by [20, 26](see [5] for a good introduction to these appraoches). The different illumination models considered in this paper belongs to that class of work. There are however others ways to cope with the change in illuminations (eg, [3, 7, 15]).

2 Overview of the hybrid tracker

The basic principle of the proposed approach is described in [22]. However, out of concern for clarity, the general overview will be briefly summed up in this section. Improvement on the texture or the illumination model will be seen in Section 3.

The approach consists of estimating the real camera pose ${}^c\mathbf{M}_o$ by minimizing the error Δ between the observed data $\mathbf{s}^*$ and the current value $\mathbf{s}$ of the same features computed using the model according to the current pose:

$$\Delta = \sum_{i=1}^{N} \rho\big(s_i(\mathbf{r}) - s_i^*\big)^2, \quad (1)$$

*e-mail: Muriel.Pressigout@irisa.fr
†e-mail: Eric.Marchand@irisa.fr

IEEE/ACM Int. Symp. on Mixed and Augmented Reality,
pp 52-55, Santa Barbara, CA, October 2006

where $\rho(u)$ is a robust function [10] introduced in the objective function in order to reduce the sensitivity to outliers (M-estimation) and $\mathbf{r}$ is a vector-based representation of the pose ${}^{c}\mathbf{M}_{o}$.

A virtual camera, defined by its position $\mathbf{r}$ in the object frame, can be virtually moved in order to minimize this error. At convergence the position of the virtual camera will be aligned with the real camera pose. This can be achieved by considering a simple control law given by $\mathbf{v} = -\lambda(\widehat{\mathbf{D}\mathbf{L}_{\mathbf{s}}})^{+}\mathbf{D}\big(\mathbf{s}(\mathbf{r})-\mathbf{s}^{*}\big)$ where $\mathbf{v}$ is the velocity screw of the virtual camera, $\mathbf{L}_{\mathbf{s}}$ is the interaction matrix or image Jacobian related to $\mathbf{s}$ and defined such as $\dot{\mathbf{s}} = \mathbf{L}_{\mathbf{s}}\mathbf{v}$ and $\mathbf{D}$ is a diagonal weighting matrix given by $\mathbf{D} = diag(w_1, \ldots, w_k)$. The weights w_i reflect the confidence in each feature and their computation is based on M-estimators and is described in [1, 2].

As presented in [22], two complementary kind of features are considered in this framework in order to improve the robustness and the accuracy of the tracking. $\mathbf{s}$ can be an edge-based feature or a texture-based one. In the case of a texture-based feature, $\mathbf{s}$ is a intensity value and its minimization comes to a SSD minimization between a reference template and the projected one according to a 2D transformation ${}^{2}tr_1$ which relies on the camera displacement:

$$\Delta = \sum_{i=1}^{N} \rho\big(I_2({}^{2}tr_1(\mathbf{p}_{1_i})) - I_1(\mathbf{p}_{1_i})\big)^2, \tag{2}$$

In [22], only piecewise planar structures were considered, projecting the points from the reference images using the homography. The use of the parallax enables to extend this approach to non-planar structures. In that latter case, a point $\mathbf{p}_1$ in image $\mathbf{I}_1$ expressed in homogeneous coordinates $\mathbf{p}_1 = ({}^{1}u, {}^{1}v, {}^{1}w)$, is transfered in image $\mathbf{I}_2$ as a point $\mathbf{p}_2$ by [9]:

$$\mathbf{p}_2 = {}^{2}tr_1(\mathbf{p}_1) = \mathbf{K}^{-1}\,{}^{2}\mathbf{H}_1\mathbf{K}\mathbf{p}_1 + \beta_1\mathbf{c}_2, \tag{3}$$

where $\mathbf{K}$ is the intrinsic camera parameters matrix, ${}^{2}\mathbf{H}_1$ is an homography (defined up to scale factor α) induced by a reference plane π that defines the transformation in meter coordinates between the images acquired by the camera at pose 1 and 2, the scalar β_1 is the parallax relative to the homography ${}^{2}\mathbf{H}_1$ and $\mathbf{c}_2 = \mathbf{K}\,{}^{2}\mathbf{t}_1$ the epipole projected onto the image 2 in pixel coordinates. β_1 depends only on parameters expressed in the camera 1 frame [9]. Once a camera displacement is generated, the homography ${}^{2}\mathbf{H}_1$ is given by ${}^{2}\mathbf{H}_1 = ({}^{2}\mathbf{R}_1 + \frac{{}^{2}\mathbf{t}_1}{{}^{1}d}{}^{1}\mathbf{n}^{\top})$ where ${}^{1}\mathbf{n}$ and ${}^{1}d$ are the normal and distance to the origin of the reference plane expressed in camera 1 frame. ${}^{2}\mathbf{R}_1$ and ${}^{2}\mathbf{t}_1$ are respectively the rotation matrix and the translation vector between the two camera frames. Spheric and piecewise planar objects have been considered as it will be shown in the result Section.

3 Robustness to illumination changes

We will now focus on the way to better take into account the changes in illumination. As it will be described in the experiments, such changes can occur at any time in the video sequence but also between the reference image and the images of the video. The basic criterion (2) to be minimized may be sensitive to such differences so the robustness of the approach has been studied, comparing several texture-based criteria corresponding to different photometric models.

Let us remember that the Jacobian matrix used to minimized the basic criterion (2) is:

$$\mathbf{L}_{I(\mathbf{p}_2)} = \frac{\partial I(\mathbf{p}_2)}{\partial \mathbf{r}} = \nabla_{\mathbf{x}}\mathbf{I}_2^{\top}(\mathbf{p}_2)\frac{\partial \mathbf{p}_2}{\partial \mathbf{r}}, \tag{4}$$

where $\nabla_{\mathbf{x}}\mathbf{I}_2(\mathbf{y})$ is the spatial image gradient of the image $\mathbf{I}_2$ at the location $\mathbf{y}$ and $\frac{\partial \mathbf{p}_2}{\partial \mathbf{r}} = \mathbf{L}_{\mathbf{p}_2}$ is the interaction matrix of an image point expressed in pixel coordinates.

The more sophisticated criteria that will be detailed are based on the intensity values in the neighborhood $\mathbf{W}$ of the considered point. They are derived from the Torrance-Sparrow [26] and the Phong [20] reflection models, with differents assumptions on the nature of the object and the illumination evolution (see [5] for a further presentation of the photometric models and their underlying assumptions).

Criterion Δ_1 : firstly the basic criterion (2) is replaced by a simplified criterion of the photometric normalization approach. It takes into account only changes in ambiant lightning:

$$\Delta_1 = \sum_{i=1}^{N} \rho\big(I_2({}^{2}tr_1(\mathbf{p}_{1_i})) - \mu_{2_i})) - (I_1(\mathbf{p}_{1_i}) - \mu_{1_i}))\big)^2, \tag{5}$$

with μ_{j_i} the intensity average value in the neighborhood $\mathbf{W}_i$ of $\mathbf{p}_{j_i}$ in image $\mathbf{I}_j$. Let note n the number of points in this neighborhood.

There are two ways to minimize (5). One can first consider that the change in illuminations is almost null between two successive images and consequently that the desired intensity values are given by $s_i^* = (I_1(\mathbf{p}_{1_i}) - \mu_{1_i}) + \mu_{2_i}$ where the $\mu_{j_i} = \frac{1}{n}\sum_{\mathbf{x}\in\mathbf{W}_i} I_2(\mathbf{x})$ are updated at each end of minimization for the next image treatment. The Jacobian matrix then remains the same as in (4) since $\mathbf{s}(\mathbf{r}) = I_2({}^{2}tr_1(\mathbf{p}_{1_i}))$.

However, μ_{2_i} depending on $\mathbf{I}_2$, it is more accurate to consider $s_i^* = I_1(\mathbf{p}_{1_i}) - \mu_{1_i}$. The Jacobian matrix must then take into account the variation of the intensity average and becomes:

$$\mathbf{L}_{s_i(r)} = \mathbf{L}_{I(\mathbf{p}_{2_i})} - \mathbf{L}_{\mu_{2_i}} \tag{6}$$

where $\mathbf{L}_{\mu_{2_i}} = \frac{1}{n}\sum_{\mathbf{x}\in\mathbf{W}_i}\mathbf{L}_{I_2(\mathbf{x})}$ is the Jacobian matrix of μ_{2_i} with respect to the pose parameters, the μ_{j_i} beeing now updated at each step of the minimization.

These two cases will be called respectively as criteria Δ_{1a} and Δ_{1b}.

Criterion Δ_2 : the basic criterion (2) is now replaced by the exact criterion of the photometric normalization approach. It is more realistic than the previous one, beeing able to mesure specular reflection whereas the first one just mesures a global illumination change:

$$\Delta_2 = \sum_{i=1}^{N} \rho\big(I_2({}^{2}tr_1(\mathbf{p}_{1_i})) - \lambda_i I_1(\mathbf{p}_{1_i}) - \eta_i\big)^2, \tag{7}$$

with $\lambda_i = \frac{\sigma_{2i}}{\sigma_{1i}}$ and $\eta_i = \mu_{2_i} - \frac{\sigma_{2i}\mu_{1i}}{\sigma_{1i}}$, σ_{j_i} beeing the standard deviations of the intensity in the neighborhood of $\mathbf{p}_{j_i}$ in image $\mathbf{I}_j$. As before, we can consider two cases. First, one can has $s_i^* = \lambda_i I_1(\mathbf{p}_{1_i}) - \eta_i$ with λ_i and η_i are updated at each end of minimization for the next image treatment. The Jacobian matrix remains therefore the same as for the basic criterion.

λ and η depending on $\mathbf{I}_2$, one can also consider $s_i^* = 0$. The Jacocian matrix is then more complicated and becomes:

$$\mathbf{L}_{s_i(r)} = \mathbf{L}_{I(\mathbf{p}_{2i})} - \mathbf{L}_{\mu_{2_i}} - \frac{\mathbf{L}_{\sigma_{2i}}}{\sigma_{1i}}(I_1(\mathbf{p}_{1_i}) - \mu_{1_i})) \tag{8}$$

where $\mathbf{L}_{\sigma_{2_i}} = \frac{1}{n^2\sigma_{2_i}}\sum_{\mathbf{x}}\big((I_2(\mathbf{x}) - \mu_{2_i})(\mathbf{L}_{I(\mathbf{x})} - \mathbf{L}_{\mu_{2_i}})\big)$ is the Jacobian matrix of σ_{2_i} with respect to the pose parameters.

These two cases will be called respectively as criteria Δ_{2a} and Δ_{2b}.

IEEE/ACM Int. Symp. on Mixed and Augmented Reality,
pp 52-55, Santa Barbara, CA, October 2006

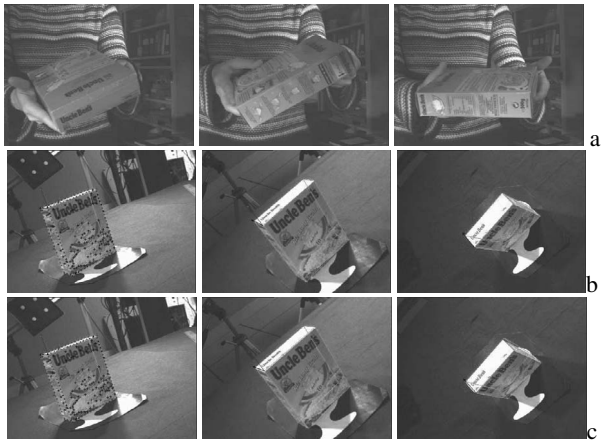

Figure 1. Experiment on the rice box. (a) some of the reference images of the texture models are show, one per face, (b) and (c) results using respectively criterion Δ_{1b} and Δ_{1a}. The tracking succeeds when using the best criterion (b), despite permanent specularities. Drift occurs at the end of (c) on the right.

4 Results

A general conclusion of this work is that the tracker is really robust to large changes in illumination and that the criteria Δ_{1a} and Δ_{1b} work better than Δ_{2a} and Δ_{2b}. As far as the comparaison between the criteria Δ_{ia} and Δ_{ib} is concerned, we found that Δ_{ib} gives better results than Δ_{ia}, $i = 1, 2$. As a result, only comparaisons between Δ_{1a} and Δ_{1b} and between Δ_{1b} and Δ_{2b} will be presented. Further discussion about those remarks will be done within the experiment description.

The first experiment points out the robustness of the tracker with respect to changes in illumination between the moment where the reference images have been captured et the moment where the object is tracked. This experiment is also representative of the difference between critera Δ_{1a} and Δ_{1b}.The second experiment shows the comparaison between criteria Δ_{1b} and Δ_{2b} and the effectiveness of the approach for AR applications.The last one is carried on a ball to demonstrate its effectivity for non-planar structures tracking in AR applications.

In all reported experiments, the edge locations and texture points used in the minimization process are displayed in the first image (blue crosses for the grey level sample locations and red crosses for the edge locations). In the next images, only the forward-projection of the model for a given pose is displayed in green.

4.1 Rice box

The objective of this experiment is to underline the robustness of the tracker with respect to changes in illumination between the reference images and the video sequence ones. The reference images compounding the texture model are given in Figure 1(a). During the experiment, the object lies on a table while the camera is moving around(see Figure 1(c)) under a different illumination. Lights are quite strong, leading to permanent specularities.

Tthe tracking remains efficient despite of these difficulties. Furthermore, one can see that there are important differences of illumination between the reference images and the video sequence: compare for examplethe first image of Figure 1(a)(the top of the rice box) with the top during the tracking in Figure 1(c). One can see the tracker is quite robust. It is not due to a small camera-object motion: the maximal motion observed on the object image between two successive images can reach suddenly 10 pixels.

This experiment illustrates also the difference between criteria Δ_{1b} (Figure 1(c)) and Δ_{1a} (Figure 1(d)). Most of time, we observe during our experiments little differences in the tracking between the two of them, using Δ_{1b} beeing more efficient. Generally, the tracker recovers from the small errors induced by Δ_{1a} but as one can see there, it may sometimes diverges. It is not a surprise that Δ_{1b} gives better results than Δ_{1a} since the Jacobian matrix takes into account the intensity average variation throughout the minimization process. The computationnal cost is of course higher but the tracking rate remains at the video acquisition rate.

4.2 CD Box

This sequence shows the difference between the criteria Δ_{1b} and Δ_{2b} in general. As it can be seen in Figure 2(a), the criterion Δ_{2b} loses the object after a while, whereas the Δ_{1b} one lasts longer. It may be surprising since Δ_{1b} is supposed to better take into account specular changes, so it should work better on such sequences with important specularities. However, the contour-based features already help the tracker to remain reliable in such cases and the instability of the standard deviations σ of the intensity values around the points in case of the saturation of the intensities or when they are homogeneous in the neighborhood may be in such cases a problem.

The latest images row in Figure 2(b) illustrates the effectiveness of the approach for AR applications. The tracking and the augmentaion has been performed at video rate.

4.3 Ball sequence

Another case where changes in illumination are omnipresent is when non-planar structures are considered. As an exemple, this section presents a ball tracking. Dealing with such a object is quite difficult as specular changes are not regular, however, the augmented images in Figure 3(b) remains consistent.

Figure 3. Ball sequence: augmented reality application. Even if the ball rotates around itself, which makes the specularities to occur step by step on the whole texture model, the tracker is effective for augmented reality applications.

5 Conclusion and perspectives

In this paper we have presented an hybrid contour and texture tracker that allows fast and robust tracking on planar and non-planar objects for augmented reality applications. The integration

Figure 2. DVD box sequence. (a) tracking result using criterion Δ_{2b}, (b) tracking result using criterion Δ_{1b}. As one can see, though criterion Δ_{2b} is supposed to better deal with specular changes, it fails sooner than the simpler criterion. (c) augmented image: the tracker is effective for such real-time applications.

of the texture-based camera motion estimation in the edge-based camera pose estimation process using the virtual visual servoing framework enables a real-time tracking requiring a CAD model and a texture model of the object. Exploiting both edge extraction and texture information to obtain a more robust and accurate pose computation. Indeed, it introduce an implicit multiple views spatio-temporal constraints in the tracking process. Considering texture (or illumination) it is important to be robust to important illumination changes. In this paper we have studied various illumination models that improve tracking robustness. Results on various planar and non-planar objects have been proposed.

References

[1] A.I. Comport, E. Marchand, M. Pressigout, F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Trans. on Visualization and Computer Graphics*, 12(4):615-628, july 2006.

[2] T. Drummond, R. Cipolla. Real-time visual tracking of complex structures. *IEEE PAMI*, 24(7):932-946, July 2002.

[3] D. Freedman, M. Turek. Illumination-invariant tracking via graph cuts. *CVPR (2)*, p. 10-17, 2005.

[4] D.B. Gennery. Visual tracking of known three-dimensional objects. *Int. J. of Computer Vision*, 7(3):243-270, 1992.

[5] M. Gouiffès, C. Collewet, C. Fernandez-Maloigne, A. Trémeau. Feature point tracking : robustness to specular highlights and lighting variations. *ECCV'2006*, LNCS 3954, p. 92-93, Graz, May 2006.

[6] M. Haag, H.H. Nagel. Combination of edge element and optical flow estimates for 3D-model-based vehicle tracking in traffic image sequences. *Int. J. of Computer Vision*, 35(3):295-319, Dec. 1999.

[7] G. Hager, P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE PAMI*, 20(10):1025-1039, October 1998.

[8] G. Hager, K. Toyama. The XVision system: A general-purpose substrate for portable real-time vision applications. *CVIU*, 69(1):23-37, Jan. 1998.

[9] R. Hartley, A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2001.

[10] P.-J. Huber. *Robust Statistics*. Wiler, New York, 1981.

[11] M. Isard, A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV,* p. 343-356, Cambridge, 1996.

[12] F. Jurie, M. Dhome. Read time 3D template matching. *CVPR*, vol. 1, p. 791-796, Hawai, Dec. 2001.

[13] C. Kervrann, F. Heitz. A hierarchical Markov modeling approach for the segmentation and tracking of deformable shapes. *Graph. Models and Image Processing*, 60(3):173-195, May 1998.

[14] H. Kollnig, H.-H. Nagel. 3D pose estimation by fitting image gradients directly to polyhedral models. *ICCV*, p. 569-574, Boston, May 1995.

[15] M. La Cascia, S. Sclaroff, V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. IEEE PAMI, 22(4):322-336, Apr 2000.

[16] D.G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE PAMI*, 13(5):441-450, May 1991.

[17] E. Marchand, P. Bouthemy, F. Chaumette, V. Moreau. Robust real-time visual tracking using a 2D-3D model-based approach. *ICCV'99*, vol. 1, p. 262-268, Kerkira, Sept 1999.

[18] E. Marchand, F. Chaumette. Virtual visual servoing: a framework for real-time augmented reality. *EUROGRAPHICS'02*, p. 289-298, Saarebrücken, Sept 2002.

[19] L. Masson, F. Jurie, M. Dhome. Contour/texture approach for visual tracking. *SCIA 2003*, p. 661-668, 2003.

[20] B.T. Phong. Illumination for computer generated pictures. *Comm. of the ACM*, 18(6):311-317, June 1975.

[21] M. Pressigout, E. Marchand. Model-free augmented reality by virtual visual servoing. *ICPR'04*, vol. 2, p. 887-891, Cambridge, Aug. 2004.

[22] M. Pressigout, E. Marchand. Real-time 3d model-based tracking: Combining edge and texture information. *IEEE ICRA'06*, p. 2726-2731, Orlando, mai 2006.

[23] J. Shi, C. Tomasi. Good features to track. *CVPR'94*, p. 593-600, Seattle, Washington, June 1994.

[24] G. Simon, M.-O. Berger. Pose estimation for planar structures. *IEEE Comp. Graphics and Applications*, 22(6):46-53, Nov. 2002.

[25] G. Taylor, L. Kleeman. Fusion of multimodal visual cues for model-based object tracking. *Australasian Conf. on Robotics and Automation*, Brisbane, Dec. 2003.

[26] K.E. Torrance, E.M. Sparrow. Theory for off-specular reflection from roughened surfacesw. *J. of the Optical Society of America*, 57:1105-1114, 1967.

[27] L. Vacchetti, V. Lepetit, P. Fua. Combining edge and texture information for real-time accurate 3d camera tracking. *ACM/IEEE ISMAR'04*, p. 48-57, Arlington, Nov. 2004.