Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid

Jean-Sébastien Franco, Edmond Boyer

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid

Jean-Sébastien Franco — Edmond Boyer

N° 5551

Avril 2005

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5551--FR+ENG

Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid

Jean-Sébastien Franco, Edmond Boyer

Thème COG — Systèmes cognitifs
Projet Movi

Rapport de recherche n° 5551 — Avril 2005 — 20 pages

Abstract: In this report, we investigate what can be inferred from several silhouette probability maps, in multi-camera environments. To this aim, we propose a new framework for multi-view silhouette cue fusion. This framework uses a space occupancy grid as a dense probabilistic 3D representation of scene contents. Such a representation is of great interest for various computer vision applications in perception, or localization for instance. Our main contribution is to introduce the occupancy grid concept, popular in the robotics community, for multi-camera environments. The idea is to consider each camera pixel as a statistical occupancy sensor. All pixel observations are then used jointly to infer where, and how likely, matter is present in the scene. As our results illustrate, this simple model has various advantages. Most sources of uncertainty are explicitly modeled, and no premature decisions about pixel labeling occur preserving therefore pixel knowledge. Consequently, scene object localization and robust volume reconstruction can be achieved, with no constraint on camera placement and object visibility. In addition, it is possible to compute improved consistent silhouettes in original views using this representation.

Key-words: computer vision, 3D Modeling from multiple views, visual hull, shape from silhouettes, occupancy grid, 3D reconstruction, sensor fusion

Fusion multi-vue d'informations de silhouettes à l'aide d'une grille d'occupation 3D

Résumé : Nous nous plaçons dans un contexte où de multiples caméras filment une scène. Nous disposons, pour chaque vue, d'images des probabilités d'appartenance à la silhouette d'objets de la scène, obtenues par un procédé de soustraction de fond. Dans ce rapport, nous explorons ce qui peut être déduit de cette information silhouette multi-vue. Dans ce but, nous proposons une nouvelle méthode pour fusionner de telles informations. Celle-ci se base sur l'utilisation d'une grille d'occupation en tant que représentation 3D probabiliste du contenu de la scène. Une telle représentation est très intéressante pour de nombreuses applications en vision, pour la perception et la localisation d'objets. Notre principale contribution est l'introduction du concept de grille d'occupation, populaire dans la communauté de robotique, dans un contexte multi-caméras. L'idée centrale est de considérer que chaque pixel est un capteur apportant de l'information sur la scène observée, et de le modéliser statistiquement comme tel. Les observations rapportées par les pixels de toutes les caméras peuvent alors être conjointement utilisées pour déduire où la matière se trouve dans la scène, et avec quelle probabilité. Comme nos résultats l'illustrent, ce modèle simple présente de nombreux avantages. La plupart des sources d'incertitude peuvent être prises en compte explicitement, et aucune décision prématurée sur l'état des pixels, et leur appartenance à une silhouette, n'est nécessaire. En conséquence, il est possible grâce à ce modèle de localiser les objets d'une scène et de recontruire leur forme, en étant robuste aux diverses sources d'incertitude. Le modèle permet de s'affranchir des contraintes de visibilité communes aux méthodes classiques de reconstruction d'enveloppes visuelles. Enfin, il est aussi possible d'utiliser l'information multi-vue acquise pour calculer des silhouettes corrigées, et cohérentes dans les vues d'origine.

Mots-clés : Vision par ordinateur, modélisation 3D à partir d'images, enveloppes visuelles, grille d'occupation, reconstruction 3D, silhouettes, fusion de capteurs

1 Introduction

Silhouette-based methods are popular for use in multi-camera environments mainly due to their simplicity and computational efficiency. These methods concern 3D modeling, multi-object localization and motion capture applications, among others. Often however in such methods, silhouettes of objects of interest are extracted using a binary labeling of pixels into foreground or background, for each view separately, and prior to any 3D operation. Unfortunately, such monocular labeling, called *background subtraction*, is difficult to achieve in a general and uncontrolled environment. Several reasons account for that, in particular perturbations due to: camera sensor noise, ambiguities between objects and background colors, changes in the lighting of the scene (including shadows of objects of interest), etc. In addition, monocular background labeling can dramatically alter 3D perception from multiple views in the presence of camera calibration errors, or if disparities between image acquisition times exist.

Our goal is therefore to find a representation of multi-view silhouette cues, where inference about silhouettes is of greater robustness to the aforementioned uncertainties than single view silhouette inference. Intuitively, the simultaneous knowledge of all images brings more information about silhouettes than knowledge from only one image. This idea has lead us to compute silhouette fusion in 3D space, in order to optimally take into account the contribution of all images. The result of such fusion naturally encodes shape information, and can therefore be used for classical modeling applications, but not only since it can also improve monocular silhouette extraction, for any silhouette-based application.

Very often silhouettes are used to infer shapes in a two step process: an individual decision about silhouette occupancy is made on a per-view basis, then shape and position are inferred geometrically from all available silhouettes using *visual hull* methods [12]. These methods can lead to a surface representation of the objects of interest [13, 15] , a voxel representation [18], or image-based representation [16]. While visual hull estimation can be exact from silhouettes [7], silhouette extraction methods come generally with several caveats resulting from the perturbations mentioned earlier. Our approach allows to delay the occupancy decision to a later stage and, as such, makes a better use of the available silhouette information.

Several methods have also been proposed to bypass silhouette estimation altogether, as many algorithms reconstruct the scene structure based only on photometric information [11]. Others possibly state it as the solution of a global optimization problem: using level sets [6], or graph cuts [8]. However all these methods have high complexities and computational costs compared to silhouette methods, as they must deal with the visibility problem for points on the surface of the object. This is why there are still many situations where silhouette methods are preferred (e.g. VR platforms, real-time setups), or used to initialize a more elaborate photometric method [10].

More closely related, Magnor *et al.* [8] propose an approach where stereo disparity and silhouettes are simultaneously estimated, with however the high computational cost of global optimization to guarantee robustness. Zeng et al. propose a multi-view background silhou-

ette extraction, based on a computationally intensive iterative geometric reasoning scheme, and with the additional constraint of common object visibility [20]. Robotics works from S. Thraun *et al.* [14] on inferring object localization from a robot-acquired image sequence are also closely related, but with significantly different contexts and assumptions. Previous papers have also explored the idea of representing the scene with a probability grid, under the different problems of wide-baseline stereo [2] or transparent objects reconstruction [1]. Grauman *et al.* [9] propose an interesting method to estimate the most probable multi-view silhouette set of humans by learning a human silhouette prior from examples, with the advantage of a higher level of semantics integration, but with limited genericity. All these approaches solve silhouette based problems in multi-camera environments with, however, limited application domains. Our approach is at a lower level, and is intended to enrich 2D silhouettes cues by embedding them into a 3D representation independently of the applications.

We propose a new framework based on the occupancy grid: a voxel grid of object occupancy probabilities in space, associated to a sensor model. The occupancy grid has been extensively used in the robotics community [5, 4], to represent a robot's environment for navigation, based on sonar and range sensor observations giving position, and eventually, orientation information. Our contribution is to extend the occupancy grid concept to image sensors, and to restate shape-from-silhouette estimation as a sensor fusion problem. To this extent, we provide each pixel with a *forward* sensor formulation, which models the pixel observation responses to the voxel occupancies in the scene. Our formulation accounts for each pixel's visibility region, voxel sampling issues, camera calibration errors, and sensor reliability. This model is in turn used to infer the answer to the more difficult inverse question: knowing the color observations, where is the matter located in the scene. We also show that the resulting occupancy grid can be used to perform multi-view background subtraction, where silhouette estimation in each view benefits from the knowledge of other views.

2 Problem Statement

We consider the problem of silhouette cue fusion from multiple views. We assume we are given a *current set* of images, obtained from fully calibrated cameras. We also assume that a set of *background images* of the scene, free from any *object of interest*, have previously been observed for each of these cameras. Importantly, no assumption is made about the existence of a visibility domain common to all cameras.

The problem is formulated as the separate Bayesian estimation, for each voxel, of how likely it is occupied by an object of interest. We formulate the problem using a forward sensor model: we model the relationship from causes to observations. Namely, in our problem, we will model how a voxel influences image formation. This enables us, using Bayesian inference, to solve the more difficult inverse problem: express the voxel occupancy likelihood using images as a noisy measurement of scene state.

Solving a Bayesian problem requires computing the joint probability of all variables of interest (which we define in section 2.1), prior to any inference. The joint probability distribution must then be decomposed and simplified, based on the main statistical dependencies we choose to consider between variables (section 3). In particular, parametric forms must be assigned to the various terms of the decomposition to explicitly model the uncertain relationship between variables (sections 3.2 and 3.3). This considerably reduces the complexity of dealing with the joint probability distribution, as it is used to infer, using Bayes' rule, the probability distribution of outputs - namely our voxel occupancies (section 4).

2.1 Main problem variables

We label the set of n current images as $\mathcal{I}$. $\mathcal{I}^i$, $i = 1 \cdots n$ is then the image data of camera i, and $\mathcal{I}^i_p$ is the image data at pixel p in image i, expressed in some color space (RGB, YUV, etc). Although not studied explicitly in this paper, additional image cues can be enclosed in the $\mathcal{I}^i_p$ term, such as the image gradient or some other local feature, without loss of generality. We assume that the image data of the corresponding m observed background images can be summarized into a single statistical model image $\mathcal{B}^i$, $i = 1 \cdots n$. Both image data sets are produced by n cameras with known projection matrices $\mathbf{P}^i$.

τ symbolizes the prior knowledge we introduce into the model. This includes what we now about the scene, what we know about sensor characteristics, our general knowledge about the system.

We define $\mathcal{G}$ as our space occupancy grid. For each space point X in the grid discretization we associate the corresponding binary occupancy variable $\mathcal{G}_X \in \{0, 1\}$, respectively free or occupied. As a common occupancy grid assumption [5], we assume statistical independence between voxel occupancies, and compute each voxel occupancy independently. In this grid, each voxel likelihood is estimated independently for tractability. Results show that independent estimation, while not as exhaustive as a global search over all voxel configurations, still provides very robust and usable information, at a much more reasonable cost.

We have defined our input and output variables. We now introduce an important hidden variable set per image, the silhouette detection maps $\mathcal{F}^i$, $i = 1 \cdots n$. These maps define, for each pixel p in image i, a binary silhouette detection variable $\mathcal{F}^i_p$. $\mathcal{F}^i_p = 1$ if the pixel sensor p in image i *reports* the presence of an object of interest anywhere along its viewing line. We insist on this definition, since there is a possibility that an object *is* indeed present along the viewing line of pixel p, but that the pixel sensor itself *fails* to detect and report this information for internal or external causes (modeling sensor failures will be discussed in section 3.2). These detection maps represent the silhouette information in our model, over which we wish to marginalize.

3 Joint Probability Decomposition

Our goal is to infer the occupancy $\mathcal{G}_X$ of a voxel at position X, given $\mathcal{I}$, $\mathcal{B}$, and τ. Thus, we must first model the impact of $\mathcal{G}_X$ on the observations. Modeling the relationships between the variables involved requires computing the joint probability of these variables, $p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)$. We propose the following decomposition, based on the statistical dependencies expressed in figure 1:

$$p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau) = p(\tau)\; p(\mathcal{B}\,|\,\tau)\; p(\mathcal{G}_X\,|\,\tau) \\ p(\mathcal{F}\,|\,\mathcal{G}_X, \tau)\; p(\mathcal{I}\,|\,\mathcal{F}, \mathcal{B}, \tau)$$

- $p(\tau)$, $p(\mathcal{B} \mid \tau)$ are the prior probabilities of our parameter set, and of background image parameters. Since we have no *a priori* reason to favor any parameter values, or background image configurations, we set these terms to a uniform distribution. They thus disappear from any subsequent inference.
- $p(\mathcal{G}_X \mid \tau)$ is the prior likelihood for occupancy, which could vary according to X for example. It is independent of all other variables except τ. As we do not wish to favor any voxel location and are mainly interested in the regularization of voxels induced by observations in this paper, we also set this term to a uniform distribution and ignore it in subsequent inferences.
- $p(\mathcal{F}\,|\,\mathcal{G}_X, \tau)$ is the silhouette likelihood term.The dependencies considered reflect that voxel occupancy in the scene explains silhouette detection in images.
- $p(\mathcal{I}\,|\,\mathcal{F}, \mathcal{B}, \tau)$ is the image likelihood term. The dependencies considered reflect that the set of image observations are only conditioned by silhouette detection in images, and the knowledge of the background color model.

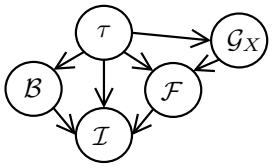

Figure 1: Variables of our system and their dependency graph. τ: prior knowledge we introduce in the model. $\mathcal{G}_X$: occupancy at voxel X. $\mathcal{B}$: background model maps. $\mathcal{F}$: silhouette detection maps. $\mathcal{I}$: observed images.

3.1 Sensor fusion simplifications

Pixel colors in input images are treated as noisy observations of the model. Like in most sensor fusion problems, we assume that the noise is independently and identically distributed.

Furthermore, the color observations at each pixel are only explained by the background image data and silhouette detection state *of this same pixel.* This induces conditional independence between all color observations $\mathcal{I}_p^i$:

$$p(\mathcal{I} \,|\, \mathcal{F}, \mathcal{B}, \tau) = \prod_{i,p} p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)$$

We also state that each silhouette detection variable itself only depends on the knowledge of the grid occupancy state, and is independent of any other sensor variable. This is also a common sensor fusion simplification: all pixel detections are considered conditionally independent, given the knowledge of their main cause, namely the voxel occupancy:

$$p(\mathcal{F} \,|\, \mathcal{G}_X, \tau) = \prod_{i,p} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau)$$

Thus, the joint probability distribution of variables of interest reduces to the following product:

$$p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau) = \prod_{i,p} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau) \tag{1}$$

We therefore have reduced the evaluation of the joint probability of all image maps, all silhouette detection maps, and our voxel occupancy, to two much friendlier subproblems. First, expressing the likelihood of silhouette detection at a single pixel, given the knowledge of occupancy of our voxel. This is the silhouette formation term (section 3.2). Second, expressing the likelihood of the color observation at a single pixel, given the silhouette detection state of this pixel, and the background color model at this pixel. This is the image formation term (section 3.3). We will now focus on these two terms.

3.2 Silhouette Formation Term

The per-pixel silhouette detection likelihood $p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau)$ models the silhouette detection response of a single pixel sensor (i, p) to the occupancy state of our voxel of interest $\mathcal{G}_X$. We need to introduce two local hidden process variables $\mathcal{S}$ and $\mathcal{R}$ to balance the influence of this voxel. Figure 2 introduces the variables and statistical dependencies of this subproblem. In an ideal and noiseless setup, the two variables $\mathcal{F}_p^i$ and $\mathcal{G}_X$ would be self-sufficient and the relationship between them expressed as simple logic: if our voxel X is occupied, and if it projects to pixel p, then silhouette detection occurs at pixel p, $\mathcal{F}_p^i = 1$. This is the implicit formulation used by all classical visual hull methods.

However, there are sources of uncertainty which perturb this intuitive reasoning. First, the assumption that a voxel lies on the viewing line of a pixel is itself uncertain. This can be due to many external causes: potential camera calibration errors, camera mis-synchronization, which both introduce misalignment in the scene. Voxel sampling is also an issue, since no voxel perfectly projects to a pixel, and its projected surface can cover several. Second, there can be causes for sensor detection other than the voxel itself: an object

occupancy other than the one related by $\mathcal{G}_X$, or a change in background scene appearance (an *internal* sensor failure due to the nature of the sensor model).

Modeling these hidden causes is possible using two boolean variables $\mathcal{S}$ and $\mathcal{R}$. This has lead us to two expressions for the silhouette detection term $p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau)$. First, let us consider the case where our voxel X is known to be occupied ($\mathcal{G}_X = 1$):

$$\begin{aligned} p(\mathcal{F}_p^i \,|\, [\mathcal{G}_X{=}1], \tau) \;&=\; p(\mathcal{S}{=}0 \,|\, \tau)\; \mathcal{U}(\mathcal{F}_p^i) \\ &+\; p(\mathcal{S}{=}1 \,|\, \tau)\; \mathcal{P}_d(\mathcal{F}_p^i) \end{aligned} \tag{2}$$

By definition, the *sampling variable* $\mathcal{S}$ equals 1 if voxel X is on the viewing line of pixel (i, p). When this is not the case ($\mathcal{S} = 0$), then the knowledge of our voxel's occupancy does not bring us any information about sensor detection, thus the uniform distribution $\mathcal{U}(\mathcal{F}_p^i)$ for silhouette detection in expression (2). If the voxel is on the viewing line of p ($\mathcal{S} = 1$), then detection by the pixel sensor is ruled by the probability distribution $\mathcal{P}_d(\mathcal{F}_p^i)$. In practice we set this distribution using a constant $P_D \in [0, 1]$, which is a parameter of our system: $\mathcal{P}_d([\mathcal{F}_p^i = 1]) = P_D$ is the detection rate of a pixel sensor, and $\mathcal{P}_d([\mathcal{F}_p^i = 0]) = 1 - P_D$ is its detection failure rate. Detection failure occurs when the pixel sensor relates that there is no matter on the viewing line, when in fact there is. This is useful for our problem: sometimes silhouette extraction fails locally. Accounting for this uncertainty in our model gives a chance to the system to still recover the correct voxel information thanks to contributions of other images.

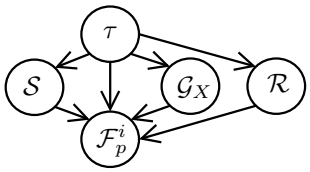

Figure 2: Variables and dependency graph of the per-pixel silhouette detection subproblem. τ: prior knowledge. $\mathcal{G}_X$: voxel occupancy. $\mathcal{S}$: sampling variable. $\mathcal{R}$: external detection cause. $\mathcal{F}_p^i$: silhouette detection at pixel (i, p).

Now, let us consider the case where our voxel is known to be empty ($\mathcal{G}_X = 0$):

$$\begin{aligned} p(\mathcal{F}_p^i \,|\, [\mathcal{G}_X{=}0], \tau) = p(\mathcal{S}{=}0 \,|\, \tau)\; &\mathcal{U}(\mathcal{F}_p^i) \\ + p(\mathcal{S}{=}1 \,|\, \tau)\; \big[\; &p(\mathcal{R}{=}1 \,|\, \tau)\; \mathcal{P}_d(\mathcal{F}_p^i) \\ + &p(\mathcal{R}{=}0 \,|\, \tau)\; \mathcal{P}_f(\mathcal{F}_p^i) \;\big] \end{aligned} \tag{3}$$

Still, no knowledge can be inferred about detection when the voxel is not on the viewing line of p ($\mathcal{S} = 0$). Yet in the case where voxel X is on the viewing line of this pixel ($\mathcal{S} = 1$), we cannot yet draw conclusions about its detection state. By definition, $\mathcal{R} = 1$ accounts for the possibility that some other object lies on the same viewing line as the voxel: in this

case detection is again ruled by the distribution $\mathcal{P}_d(\mathcal{F}_p^i)$. However, in the case no other object obstructs the viewing line ($\mathcal{R} = 0$), detection is ruled by distribution $\mathcal{P}_f(\mathcal{F}_p^i)$. We set this distribution using a constant $P_{FA} \in [0, 1]$, which is a parameter of our system: $\mathcal{P}_f([\mathcal{F}_p^i = 1]) = P_{FA}$ is the false alarm rate of a pixel sensor. It is the rate with which the sensor falsely relates the presence of matter on its viewing line when in fact there is none. $\mathcal{P}_f([\mathcal{F}_p^i = 0]) = 1 - P_{FA}$ is the expected rate with which we expect this pixel to correctly report non-detection.

We must assign a parametric form to $p(\mathcal{R} \,|\, \tau)$. There can be detection causes anywhere along the viewing line of p. We make no assumption about these causes and consider that detection is equally likely to be triggered by the voxel occupancy or by these causes. We therefore set this term to uniform.

Parametric form for Sampling Term $p(\mathcal{S} \,|\, \tau)$. This term is dependent on i, p and X. We use uniform sampling, with $p(\mathcal{S} \,|\, \tau) = \mathcal{U}_{k\times k}(x - p)$. This gives equal weight to all voxels that fall within a $k \times k$ window around pixel p. A smoother, normal-based sampling could also be used but requires a higher computational cost to integrate information. Generally, the shape of this sampling function can easily be modified for specific needs.

Both uniform and normal sampling forms enable some control over calibration, mis-synchronization, and some classification errors: several pixels will be able to contribute to a single voxel's decision upon inference. Thanks to the introduction of these two hidden processes and the given parametric forms, our method unifies broad silhouette uncertainty management and simple image sampling methods used in some visual hull algorithms such as [3].

3.3 Image Formation Term

The image pixel likelihood term $p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)$ seeks to explain the color information of a pixel (i, p), given the knowledge of the background color model at this pixel and whether or not a silhouette detection occurred at this pixel. We now explain what parametric forms we give to this term.

First, let us consider the case where a silhouette detection occurred at pixel (i, p). The knowledge about background images does not bring us any more information about the pixel's expected observed color, because we know that the background object is occluded by an object of interest, whose color the pixel observes. As we make no assumptions about the color of objects of interest, we set the posterior distribution of observed colors to uniform in this case:

$$p(\mathcal{I}_p^i \,|\, [\mathcal{F}_p^i = 1], \mathcal{B}_p^i, \tau) = \mathcal{U}(\mathcal{I}_p^i)$$

The second case we need to consider is the case where no silhouette detection has occurred at this pixel. Intuitively, if the sensor is known to report that there are no objects of interest on its viewing line, then the observed color at this pixel should look similar to the background color. In practice, we choose to summarize all background images by estimating the parameters $\mathcal{B}_p^i = (\mu_p^i, \sigma_p^i)$ of a normal distribution in (Y,U,V) space for each pixel. Our

expectancies about the observed color at pixel (i,p) can therefore be formulated naturally with this normal distribution:

$$p(\mathcal{I}_p^i \,|\, [\mathcal{F}_p^i = 0], [\mathcal{B}_p^i = (\mu_p^i, \sigma_p^i)], \tau) = \mathcal{N}(\mathcal{I}_p^i \,|\, \mu_p^i, \sigma_p^i)$$

This background color representation derives from classical background subtraction methods [19]. Importantly however, the framework we present is independent of the chosen background model, and could easily account for any other background modeling technique such as a mixture of Gaussian distributions [17], which better accounts for sub-pixel ambiguities in background image formation. Nevertheless, some problems persist whatever the background model: color ambiguities between foreground and background objects, light change or scene geometry change. It is the goal of our integrated multi-view approach to compensate for the weaknesses of single-view estimation.

4 Voxel Occupancy Inference

Once the joint probability distribution has been fully determined, it is possible to use Bayes' rule to infer the probability distributions of our *searched* variable $\mathcal{G}_X$, given the value of our *known* variables $\mathcal{I}, \mathcal{B}, \tau$, and marginalizing over *unknown* variables $\mathcal{F}$:

$$\begin{aligned} p(\mathcal{G}_X \,|\, \mathcal{I}, \mathcal{B}, \tau) &= \frac{\sum_{\mathcal{F}} p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)}{\sum_{\mathcal{G}_X, \mathcal{F}} p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)} \\ &= \frac{\sum_{\mathcal{F}} \prod_{i,p} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)}{\sum_{\mathcal{G}_X, \mathcal{F}} \prod_{i,p} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)} \qquad (4) \\ &= \frac{\prod_{i,p} \sum_{\mathcal{F}_p^i} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)}{\sum_{\mathcal{G}_X} \prod_{i,p} \sum_{\mathcal{F}_p^i} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)} \qquad (5) \end{aligned}$$

where (1) has been substituted in (4). This inference is simplified thanks to the fact that all sums on $\mathcal{F}_p^i$ variables can be performed at the pixel level and thus factorized (5). Namely, $\mathcal{F}_p^i$ can itself be considered a local pixel process which balances a generic sensor model term $p(\mathcal{I}_p^i \mid \mathcal{G}_X, \mathcal{B}_p^i, \tau)$. This term directly relates the pixel's color observation to voxel occupancy, using the image formation and silhouette formation terms previously discussed, and marginalizing over silhouette detection $\mathcal{F}_p^i$, as expressed below:

$$p(\mathcal{I}_p^i \,|\, \mathcal{G}_X, \mathcal{B}_p^i, \tau) = \sum_{\mathcal{F}_p^i} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)$$

This sensor term clarifies the inference (5), as it shows how each pixel contributes to the voxel occupancy probability:

$$p(\mathcal{G}_X \,|\, \mathcal{I}, \mathcal{B}, \tau) = \frac{\prod_{i,p} p(\mathcal{I}_p^i \,|\, \mathcal{G}_X, \mathcal{B}_p^i, \tau)}{\sum_{\mathcal{G}_X} \prod_{i,p} p(\mathcal{I}_p^i \,|\, \mathcal{G}_X, \mathcal{B}_p^i, \tau)} \qquad (6)$$

Note that the final inference expression (6) deceptively relates our voxel occupancy to *all* pixel observations, which is of course intractable, given that we must perform this inference *for each voxel* of the grid. In practice, the sampling schemes discussed in section 3.2 bound the region of influence of a pixel in the images. Detection probabilities of pixels too far from the voxel's projection degenerate to uniform, as expressed in equations (2) and (3). Their contribution therefore factors out of the inference expression (6). This makes the inference tractable, by computing this product over a local window of pixels, centered at the image projection of X, in each image. Note that such a product quickly reaches machine precision limits: in practice we compute the inference using sums of log probabilities. If k is the size of the window, and N the number of voxels per dimension, then the complexity of inferring all voxels of the grid is $O(n\,k^2N^3)$.

5 Results and Applications

We have implemented this silhouette cue fusion algorithm, using uniform voxel sampling for experiments. Compared to normal sampling it is a good trade-off between computational cost and power of information integration. Note that the method has only three parameters $\{P_D, P_{FA}, k\}$, respectively the detection and false alarm rates, and the sampling window size. Often these parameters can be fixed once and for all for a given application. P_D and P_{FA} ponderate the confidence given to the observations. If $P_{FA} = 0$ and $P_D = 1$, then we trust observations blindly. If P_{FA} and P_D are close to 0.5 then observations are not trustworthy: it takes many more observations to conclude about the occupancy. k decides how many observations we locally consider in each image and therefore also ponderates the decision. We have tested the algorithm under various conditions, as it can be applied to many application fields.

5.1 Modeling from Images

The grid itself is an estimate of shape. We illustrate this using the walking sequence. This sequence of a person walking inside the test room was acquired using 8 cameras of different resolutions and characteristics (640x480, 780x580) at 15Hz. As figure 3 illustrates, the silhouette information that can be retrieved using monocular background subtraction is noisy. Also note that some cameras may not see the object in its entirety during the sequence, as is the case in the second image here, where the person's left forearm is out of sight. The color model used for these single-view subtractions is the same as in our model (Gaussian distribution for background colors). These subtractions summarize the silhouette information available to our algorithm in each image.

Figure 4 shows the results of our method on a frame of the walking sequence, using a 120^3 grid. Vertical and horizontal cross-sections of the grid are supplied to give an intuition of what information is available in the grid. A more dynamic view is given in the supplemental

Figure 3: Inputs. (a) Four of the eight input images of the walking sequence (8 cameras, 15 images/sec acquisition) (b). The corresponding result given by monocular subtraction using the same background color model as in our method (semi-transparent rendering pondered by silhouette probability). Note the imperfections in these silhouettes: camera 2 misses the subject's left arm. There are holes in the silhouette in various images.

cross-sections video [1]. As shown in figure 4(c), good surface modeling results can be achieved by extracting an isosurface from the probability grid. Fine detail of the surface is recovered, and holes occurring in monocular subtractions are often filled. Additional modeling results are shown in figure 5 and in the supplied video[2].

The classical voxel-based visual hull approach has been implemented for comparison, and results are shown in figure 4(d), where each voxel is projected in images and carved if it is outside silhouettes. We use the background subtractions of figure 3 for this experiment, and manually choose the best threshold *in each image* to provide binary silhouettes to the algorithm. Some holes are left unfilled by this method. Note that our method recovers valid occupancies from views that don't see the object entirely. This is transparent to the algorithm, because it only integrates information from sensors which see the voxels. This is unlike all classical, surface or volumetric visual hull approaches, where explicit assumptions must be made about voxels that project outside the visibility domain of an image 4(d)).

[1] http://www.inrialpes.fr/movi/pub/video/cross-section.avi
[2] http://www.inrialpes.fr/movi/pub/video/sequence.avi

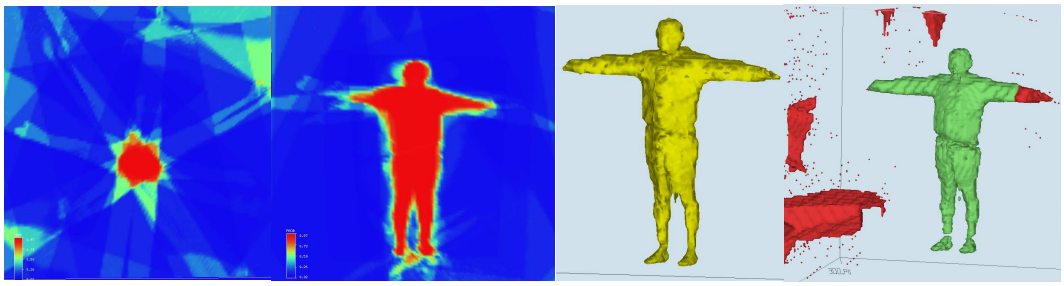

Figure 4: A snapshot of the walking sequence. Acquisition 15imgs/sec, 8 cameras, using inputs of figure 3. Voxel grid 120^3. Computation time: approx 8 sec on a 2.4 GHz PC. Parameters used: $P_D = 0.9, P_{FA} = 0.1, k = 5$ (a) Horizontal (chest) cross-section of the probability grid. Greenish regions on the upper right are not seen by any camera (probability 0.5). (b) vertical grid cross-section. (c) Isosurface of probability 0.80 obtained from the grid. Shadow noise in images does not perturb the estimation, holes are filled. (d) Two classical visual hull reconstruction schemes: in green, assuming common visibility of the object by all cameras. The arm is lost. In red, assuming that what is outside the visibility domain of a camera can be part of the visual hull. The former recovers the left arm, but ghost objects appear.

5.2 Multi-View Background Subtraction

Our method computes a fusion of silhouette cues. This information can be used to compute consistent silhouettes in our input images, by re-projecting and rendering the occupancy grid from our input views, using a maximum intensity projection approach (see figure 6). This heuristic approximates computing the inference of pixel silhouette state given the knowledge of all image observations, which is possible but expensive given our statistical model. The maximum intensity projection of occupancy probabilities enables to express where silhouette detection is most likely in images at a much more reasonable cost.

This re-rendering heuristic defines a multi-view background subtraction procedure. In particular, a single threshold can be chosen for all images simultaneously to find optimal binary silhouettes, as shown in the figure. Fine detail is preserved, being only limited by grid resolution. Each view benefits from the knowledge of silhouette information in the other views.

5.3 Object detection

The method can be used in much harder conditions to infer information about a scene. In particular, in the presence of high levels of noise, the size of the sampling window can be increased for additional robustness, with however a negative impact on precision (as this operation tends to dilate the probability volume, and the underlying isosurfaces consequently).

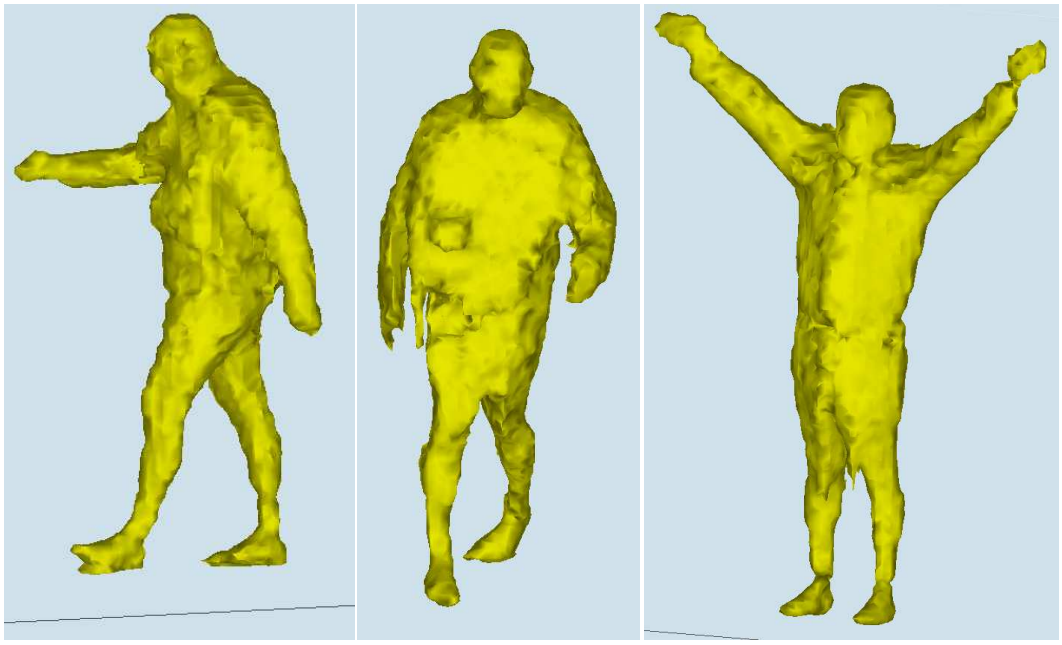

Figure 5: Isosurface of probability 0.80 at different time instants of the walking sequence. See video.

Very noisy conditions limit the use of the method for 3D modeling; however the method can still be used reliably to *locate* objects in the scene - without necessarily reconstructing their precise surface. We illustrate the potential use of our method for object localization in a scene, with loose camera configurations and poor contrast images, in figure 7. In this experiment, 8 cameras are placed such that a relatively large area ($25m^2$) can be monitored in the room. Only the center of the room is seen by a majority of cameras, the peripheral regions of this area are seen by 3 or 4 cameras at most. A video of the experimental results is available[3]. Two people walk randomly in the scene and are successfully localized, when seen by at least three cameras. It is an interesting empirical result: two cameras aren't enough to distinguish the object from equally probable ghosts objects arising from visual ambiguities. These visual ambiguity regions are empty regions of space which are usually in the shadow of a real object (and as such inside a silhouette cone), and where no other camera correctly reports the emptiness of the region. They can be seen observed in the results video.

[3]http://www.inrialpes.fr/movi/pub/video/localization.avi

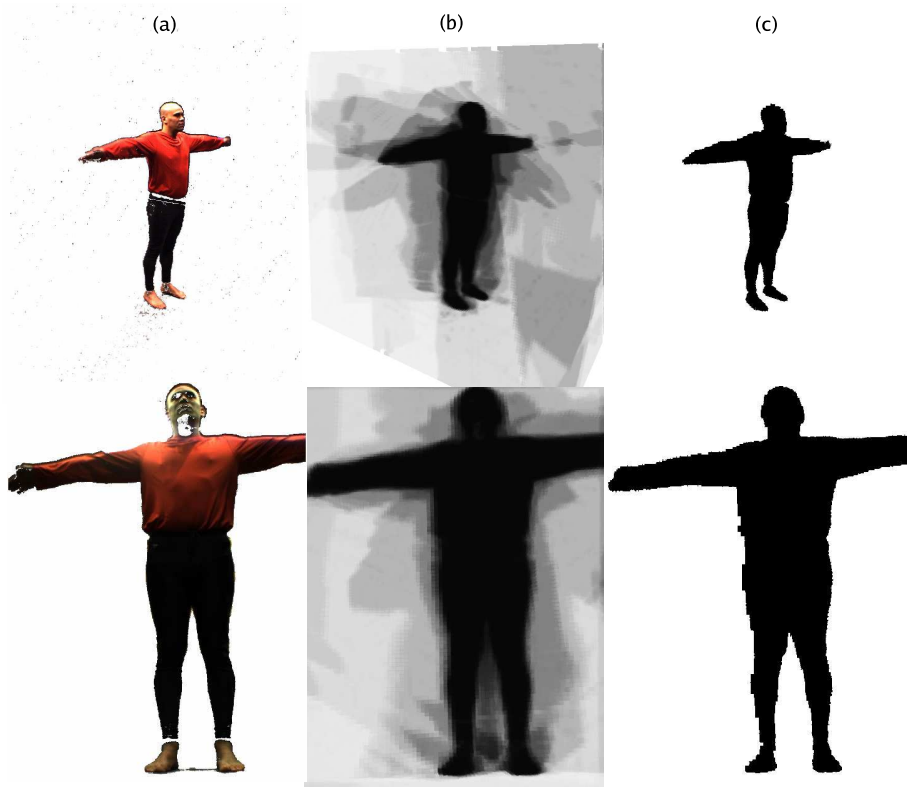

Figure 6: Two examples of silhouette re-rendering. (a) Monocular subtraction (semi-transparent rendering weighted by silhouette probability). Note the imperfections: in the top image, the subject is cut in half at the waist; in the bottom image, both feet are separated from legs and the face has huge holes. No monocular repairing schemes such as morphological operations can recover these artifacts *and* give decent precisions for recovered silhouettes. (b) Maximum intensity projection rendering of occupancy grid (120^3) probabilities from original viewpoints. (c) The same data after an optimal threshold was manually chosen *for all silhouettes simultaneously*: silhouettes are enhanced. Some aliasing artifacts may appear depending on grid resolution and scene configuration.

6 Discussion

We have presented a novel approach for silhouette cue fusion from multiple views. We use a rigorous sensor fusion framework, to relate scene information directly to observations. This has various advantages: the entire chain from causes to observations is modeled and all assumptions made explicit. It also avoids making hard decisions about silhouette labeling in images, which would require per-image parameter settings. Thus the underlying silhouette information in images can be smoothly integrated, using only three global parameters of

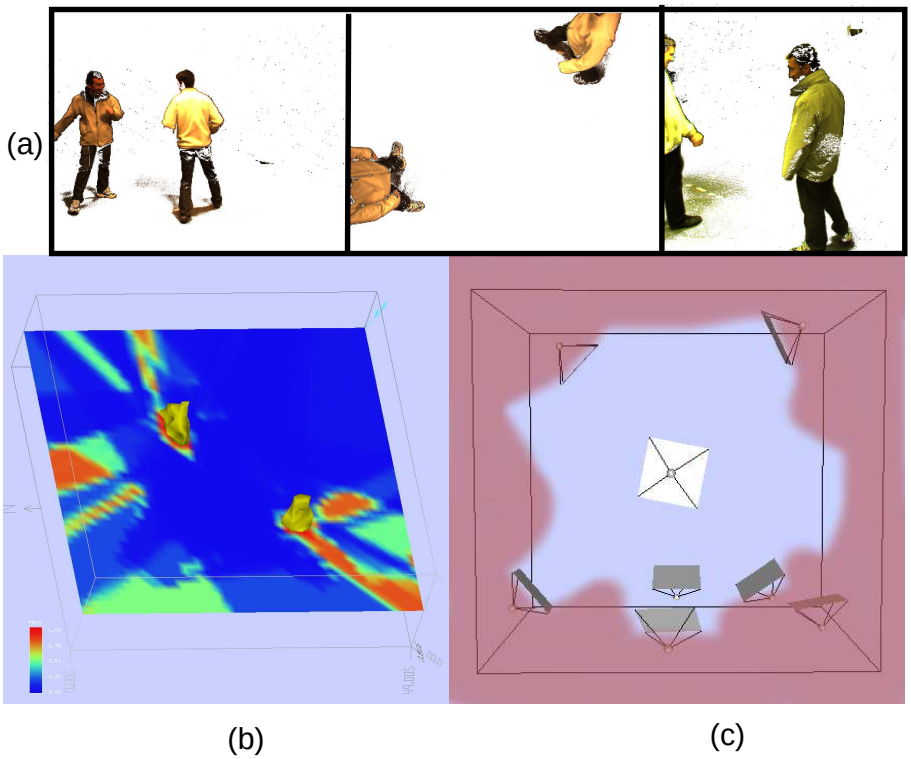

Figure 7: Multi-object sequence, with 8 cameras. (a) Monocular subtraction of some of the input views (semi-transparent rendering weighted by silhouette probability). Poor lighting and contrast create difficult conditions for single-view subtraction. (b) Our method used to reconstruct a coarse grid (50x50x18) of the scene, sufficient to localize objects (using $k = 25$). Computation time: 7s/frame. A horizontal cross section of occupancy probabilities is given, as well as 0.67 probability isosurfaces showing localized objects (see video). (c) Camera configuration in this scene. Dark-reddish regions are less reliable as they are seen by 2 cameras or less. Nevertheless the system is able to detect the presence of objects in a 5x5 meter region.

a pixel sensor model. These parameters intuitively control the reliability of observations. This approach has been validated with several applications, and many new ideas can be experimented and plugged-in without changing the core of the method.

Arguably, more dependencies could be considered in the model. Namely, we notice that reliability of pixel decision can be related to the colors observed at this pixel. For example we observe many times the case where black foreground objects are observed in front of a black background, which could call for special treatment and will be investigated. More generally, our model estimates static grids at one time instant. It would greatly benefit

from temporal consistency, where passed observations are used to infer current occupancy states. Happily, occupancy grids provide a good framework for temporal accumulation of information, being one of its main uses in the robotics community [4]. We will investigate these possibilities to extend the capabilities of our system.

References

[1] J. S. D. Bonet and P. A. Viola. Roxels: Responsibility weighted 3d volume reconstruction. In *Proceedings of the 7th International Conference on Computer Vision, Kerkyra, (Greece)*, volume I, pages 418–425, Sept. 1999.

[2] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, (Canada)*, volume I, pages 388–393, 2001.

[3] G. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, (USA)*, volume II, pages 714 – 720, June 2000.

[4] C. Coue. *Modèle bayésien pour l'analyse multimodale d'environnements dynamiques et encombrés : application à l'assistance à la consuite automobile en milieu urbain.* PhD thesis, Institut National Polytechnique de Grenoble, Dec. 2003.

[5] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *IEEE Computer, Special Issue on Autonomous Intelligent Machines*, 22(6):46–57, June 1989.

[6] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proceedings, 5th European Conference on Computer Vision, Freiburg, (Germany)*, volume I of *Lecture Notes in Computer Science*, pages 379–393. Springer, June 1998.

[7] J.-S. Franco and E. Boyer. Exact Polyhedral Visual Hulls. In *Proceedings of the British Machine Vision Conference, Norwich (UK)*, pages 329–338, Sept. 2003.

[8] B. Goldlücke and M. Magnor. Joint 3-d reconstruction and background separation in multiple views using graph cuts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, (USA)*, volume I, pages 683–694, June 2003.

[9] K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, (USA)*, volume I, pages 187–194, June 2003.

[10] J. Isidoro and S. Sclaroff. Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. In *Proceedings of the 9th International Conference on Computer Vision, Nice, (France)*, pages 1335–1342, 2003.

[11] K. Kutulakos and S. Seitz. A Theory of Shape by Space Carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[12] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *IEEE Transactions on PAMI*, 16(2):150–162, Feb. 1994.

[13] S. Lazebnik, E. Boyer, and J. Ponce. On How to Compute Exact Visual Hulls of Object Bounded by Smooth Surfaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Kauai, (USA)*, volume I, pages 156–161, December 2001.

[14] D. Margaritis and S. Thrun. Learning to locate an object in 3d space from a sequence of camera images. In *International Conference on Machine Learning*, pages 332–340, 1998.

[15] W. Matusik, C. Buehler, and L. McMillan. Polyhedral Visual Hulls for Real-Time Rendering. In *Eurographics Workshop on Rendering*, 2001.

[16] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image Based Visual Hulls. In *ACM Computer Graphics (Proceedings Siggraph)*, pages 369–374, 2000.

[17] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, (USA)*, volume II, pages 246–252, June 1999.

[18] R. Szeliski. Rapid Octree Construction from Image Sequences. *Computer Vision, Graphics and Image Processing*, 58(1):23–32, 1993.

[19] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[20] G. Zeng and L. Quan. Silhouette extraction from multiple images of an unknown background. In *Proceedings of the 6th Asian Conference on Computer Vision, Jeju Island, (Korea)*, Jan. 2004.

Contents

1 Introduction **3**

2 Problem Statement **4**
2.1 Main problem variables 5

3 Joint Probability Decomposition **6**
3.1 Sensor fusion simplifications 6
3.2 Silhouette Formation Term 7
3.3 Image Formation Term 9

4 Voxel Occupancy Inference **10**

5 Results and Applications **11**
5.1 Modeling from Images 11
5.2 Multi-View Background Subtraction 13
5.3 Object detection 13

6 Discussion **15**

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399