

**Traitements Vidéo  
et  
Espaces Virtuels de Réunion**

J.-L. DUGELAY & S. VALENTE

EURECOM, MultiMedia Communications dept.  
2229, route des Crêtes, BP 193, F - 06904 Sophia Antipolis Cedex  
url. <http://www.eurecom.fr/~image>  
e-mail. [dugelay@eurecom.fr](mailto:dugelay@eurecom.fr)  
tel. +33 93 00 26 41  
fax.+33 93 00 26 27

**Résumé** : Cet article présente les techniques de traitements vidéo développées dans le cadre d'un projet de télé-virtualité (projet "TRAIVI"). Ce projet vise à mettre en place des espaces virtuels de téléconférence via des liaisons bas-débit. Sont présentés dans ce papier, les premiers résultats obtenus en clonage et en spatialisation vidéo. Le clonage de visages est réalisé par analyse d'images et utilise un modèle CYBERWARE du locuteur pour la restitution. La spatialisation vidéo permet de reconstruire, par interpolation à partir d'un triplet de vues de référence, des points de vue fictifs de l'espace de réunion.

**mots clés** : télé-virtualité, communications vidéo, clonage, spatialisation vidéo.

**Video Processing  
and  
Virtual Meeting Rooms**

**Abstract** : This paper describes video processing techniques developed in the context of a tele-virtuality project (project "TRAIVI"). This project intends to create and run virtual meeting places via low bit rate bindings. In this paper, preliminary results in cloning and video spatialization are described. The cloning of faces is realized by image processing and using a CYBERWARE model of the speaker for displaying. The video spatialization allows the reconstruction of some virtual points of view of the meeting room, by interpolating a set of three reference views.

**Keywords** : tele-virtuality, video communications, cloning, video spatialization.

## 1. Introduction

*TRAVI* est un projet de télé-virtualité. Ce projet vise à mettre en place des espaces virtuels de téléconférence via des liaisons bas débit. Il s'agit d'implanter une structure de communication audio-vidéo entre plusieurs sites distants afin que différentes personnes puissent converser confortablement, en réduisant au maximum l'impression de distance grâce à des outils de la réalité virtuelle. Pour atteindre un niveau de réalisme audiovisuel satisfaisant dans la création et la gestion de ces espaces virtuels de réunion, plusieurs techniques de traitement audio et vidéo doivent être parfaitement maîtrisées. Ces techniques sont la spatialisation et le multiplexage audio, annulation d'écho, la spatialisation vidéo, le clonage des participants, la synchronisation audio-vidéo,... Ce papier traite de l'aspect vidéo, et plus particulièrement de la spatialisation et du clonage.

Actuellement, une réunion audio-vidéo à plusieurs participants peut se tenir à distance via des liaisons et des équipements spécialisés. Ces systèmes offrent une qualité de service acceptable pour des réunions limitées à deux participants, ou plus généralement à deux sites. Au-delà de deux, l'environnement proposés par les systèmes actuels devient critique et les communications entre participants deviennent beaucoup plus difficiles qu'elles ne le sont dans le cas d'une réunion réelle. Un système de vignette de chaque participant et/ou un affichage alterné des différents sites n'offrent pas, à ce jour, un service satisfaisant.

Le modèle de visage de chaque participant ainsi que plusieurs vues de référence de la salle de réunion sont supposés être soit déjà disponibles sur les sites récepteurs, soit préalablement télé-chargés en début de réunion. Afin de limiter la quantité d'informations à transmettre, seuls les paramètres de mise à jour du clone de chaque participant ainsi que ceux caractérisant, à chaque instant, le point de vue à afficher de l'espace de réunion seront transmis. Le point de vue à restituer sur chaque site ne sera pas identique. En effet, ce point de vue doit être cohérent par rapport à la position de chaque participant autour de la table de réunion et de son centre d'intérêt dans la scène. Il est à noter qu'en conséquence une partie des informations vidéo transmises vers chaque site participant sera spécifique.

Ce papier présente les premiers résultats obtenus en clonage et en spatialisation vidéo. Le clonage de visages est réalisé par analyse d'images et utilise un modèle CYBERWARE de chaque participant pour la restitution (section II). La spatialisation vidéo permet de reconstruire, par interpolation à partir d'un triplet de vues de référence, des points de vue fictifs de l'espace de réunion (section III). Des traitements similaires à ceux présentés en vidéo seront également à considérer en audio [1].

## 2. Clonage

Le clonage est l'une des techniques clé pour la création d'espaces virtuels de communication et de télé-présence [2].

Le clonage de visage, pour cette application, utilise un modèle CYBERWARE pour la restitution afin de garantir un niveau acceptable de réalisme. Un modèle CYBERWARE est obtenu par le biais d'une acquisition tridimensionnelle d'un visage réel. Ce modèle est donc propre à une seule personne. Les données sont matérialisées par deux fichiers numériques: le premier contient un ensemble de points 3D représentant la géométrie de la tête, et le second contient les informations de texture associées à ces points (illustration n°1).

Le télé-asservissement d'un modèle CYBERWARE est ici divisé en deux parties: globale et locale. La détection, puis la restitution, des mouvements globaux de la tête sont décrits dans la section 2.1. Différentes approches sont envisagées, dans la section 2.2, pour

assurer la restitution des expressions faciales, incluant le mouvement de la bouche, des yeux, des paupières, des sourcils,...



**illustration n°1: Textures associées au modèle CYBERWARE de "heidi"**

## 2.1 Animation globale

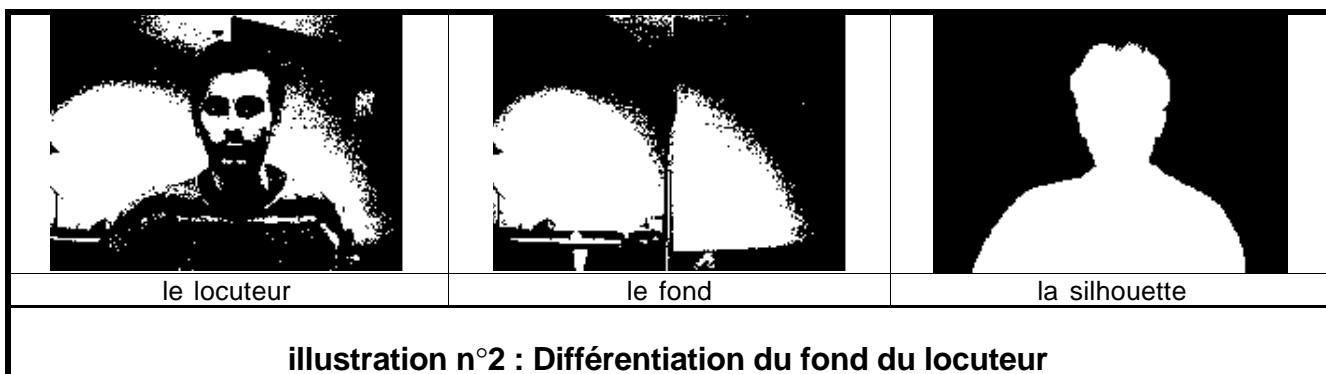
Afin de valider les techniques développées en analyse et synthèse d'images, le schéma suivant de télé-asservissement global d'un modèle CYBERWARE, incluant la transmission des paramètres, a été simulé [3] :

- . un individu  $\lambda$ , situé devant sa station de travail sur un site 1 (dit site émetteur), est filmé par une caméra,
- . les images acquises sont analysées afin d'extraire des informations sur la position de la tête;
- . les informations ainsi extraites sont transmises au site distant 2 (dit site récepteur);
- . le site 2, qui possède localement en base de données le modèle "CYBERWARE" de l'individu  $\lambda$ , interprète les paramètres transmis et restitue en conséquence le visage cloné de  $\lambda$ .

### 2.1.1 Analyse

La première étape consiste à séparer la silhouette du locuteur du fond de la scène. Cette opération est réalisée simplement par une différence seuillée entre l'image courante et une image de référence de la scène sans locuteur (illustration n°2). Cette approche requiert que le fond de la scène soit statique.

Le visage est à chaque instant englobé par une fenêtre rectangulaire [4]. Au sein de cette fenêtre l'axe médian du visage ainsi que l'axe des yeux sont détectés (illustration n°3).



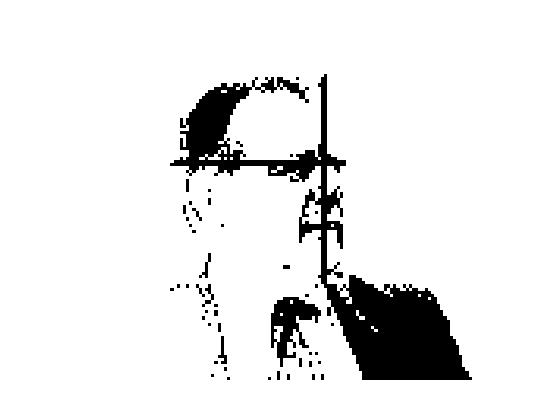



## 2.1.2 Transmission

La taille de la fenêtre, ainsi que les coordonnées associées respectivement au centre de la fenêtre d'une part et à l'intersection entre les deux axes précédemment définis d'autre part sont transmis via le réseau en utilisant les sockets UNIX du protocole TCP/IP.

## 2.1.3 Synthèse

Les paramètres 2D issus de l'analyse d'images sont reçus puis interprétés en termes 3D : la position du centre de la fenêtre permet de déterminer deux des trois degrés de liberté en translation. Le troisième degré de translation, qui traduit un rapprochement ou un éloignement du conférencier devant sa station, est calculé à partir de la taille de la fenêtre. Deux des trois degrés de liberté en rotation (vers le haut/bas et vers la droite/gauche) sont estimés à partir des coordonnées de l'intersection de l'axe des yeux avec l'axe médian du visage.

	
le locuteur est devant sa station de travail. La position de sa tête, puis de ses yeux sont automatiquement détectés par analyse d'images.	le modèle CYBERWARE du locuteur est affiché sur un site distant.
	
le locuteur tourne la tête vers sa gauche.	le modèle CYBERWARE du locuteur tourne également à gauche.

### illustration n°3 : Télé-asservissement d'un modèle CYBERWARE (\*) (\*\*) (\*\*\*)

(\*) Cette simulation tourne actuellement, en temps réel, entre une station SUN Sparc ZX (site émetteur) et une station SGI Impact (site récepteur). L'analyse (resp. la synthèse) utilise certaines routines de la bibliothèque `Xil` (resp. `OpenGl`).

(\*\*) Sur cet exemple, le locuteur utilise son propre modèle CYBERWARE mais il peut aisément utiliser celui d'une tierce personne.

(\*\*\*) Les signaux audio sont provisoirement transmis sans traitement particulier. Les images n'étant plus transmises explicitement puisque seuls quelques paramètres de mise à jour du clone le sont, des problèmes de synchronisation audio/vidéo peuvent apparaître.

Le paragraphe suivant aborde la partie la plus délicate en clonage de visage, à savoir la détection puis la restitution des expressions faciales des participants.

## 2.2 Animation locale

Les travaux actuels en clonage de visage utilisent le plus souvent des marqueurs (rouge à lèvres, pastilles,...) disposés sur le visage à cloner. Dans le cadre de cette étude, il s'agira de s'affranchir de cette contrainte par analyse d'images, éventuellement complétée par une analyse des signaux audio.

Plusieurs stratégies, éventuellement complémentaires, sont envisagées pour détecter puis animer localement le visage des participants.

Certaines expressions telles que le mouvement des yeux et des sourcils peuvent être simulées lors de l'affichage par la seule modification des informations de texture associées aux données géométriques du modèle CYBERWARE. A chaque modèle géométrique, il sera associé, non plus un seul fichier de texture mais plusieurs. Ainsi, à chaque instant, l'un des fichiers de texture sera sélectionné à la restitution, en fonction de l'état du locuteur identifié lors de l'étape d'analyse (les yeux vers la gauche ou vers la droite, la bouche ouverte ou fermée,...). Pour d'autres expressions, il sera de plus nécessaire de compléter la texture du modèle par des portions d'images réelles. Typiquement, si la langue ou les dents d'un participant sont visibles lors de sa locution alors que ces éléments ne figuraient pas dans le modèle de référence, les portions d'images réelles correspondantes devront être segmentées lors de l'étape d'analyse, puis transmises, et enfin mises à l'échelle pour être incrustées dans la texture artificielle du modèle CYBERWARE.

La vidéo et l'audio peuvent être tous deux utilisés pour détecter le mouvement de la bouche. Il est possible par analyse d'images, à l'aide de contours actifs, d'évaluer avec précision l'état spatio-temporel des lèvres du locuteur (aire, déformations, périmètres) [5]. Ces informations extraites, il s'agira ensuite de les répercuter sur le modèle lors de l'affichage. Pour se faire, l'ensemble des coordonnées 3D représentant la partie géométrique du modèle CYBERWARE sera manipulé à l'aide d'un maillage actif, sur lequel un modèle de déformation du visage pourra ensuite être appliqué [6]. A défaut de restituer les expressions réelles des participants, il sera également tout à fait acceptable, dans le cadre de cette application, d'afficher des expressions, uniquement réalistes et non réelles, générées par exemple à partir des signaux de parole [7].

Les informations nécessaires à l'animation faciale seront à ajouter à celles déjà transmises pour l'asservissement global du modèle CYBERWARE. Néanmoins, la quantité globale d'informations à transmettre restera extrêmement faible au regard du volume initial d'informations. Ce schéma de clonage est sur cet point proche d'autres schémas développés en codage orienté objet pour des applications en visiophonie à très bas-débit où les séquences traitées sont des scènes de type "head & shoulders" [8]. Une étape supplémentaire de compression peut se faire en utilisant un dictionnaire prédéfini d'expressions faciales. Il suffira alors de ne transmettre qu'une suite d'index (choisis par analyse au niveau du site émetteur) pour animer, de manière réaliste mais pas obligatoirement exacte, les visages au niveau des sites récepteurs.

Ces clones devront être ensuite intégrés dans l'espace virtuel de réunion. Cette étape réalisée, chaque participant devra visualiser l'espace de réunion sous un angle cohérent par rapport à sa place virtuelle (chacun voyant les autres sauf lui-même) mais également par rapport à la direction de son regard, préalablement détectée lors du clonage de son visage.

### 3. Spatialisation vidéo

La spatialisation vidéo a pour objectif de construire un espace complet de réunion à partir de quelques vues de référence, puis de restituer à chaque participant les images en fonction de sa position dans la réunion. C'est précisément ce qui nous différencie des systèmes de visioconférence actuels, qui imposent encore à chaque conférencier un point de vue général sur la scène. Notons qu'aucun modèle artificiel de la salle de réunion n'est créé et qu'au contraire, afin de garantir le réalisme de la scène visualisée, les resynthèses de points de vue de la scène sont produites à partir d'images réelles. Actuellement, la reconstruction d'une vue fictive est réalisée à partir de trois vues de référence. Cette partie est présentée dans la section 3.1. Des extensions sont envisagées, dans la section 3.2, afin de travailler sur un nombre de vues de référence plus important permettant une couverture visuelle complète et précise d'une salle de réunion.

#### 3.1 A partir de trois vues de référence

Les travaux décrits dans ce paragraphe s'appuient sur des publications récentes [9], et constituent une première étape vers l'objectif précédemment défini. Ces travaux, qui ont permis d'établir des relations trilineaires entre trois vues d'une même scène, permettent de reconstruire l'une des trois vues à partir des deux autres et de quelques paramètres. Bien que cette approche ne nécessite pas de calibration 3D explicite, les paramètres en question peuvent être, en première approximation, assimilés à des paramètres de calibration.

L'implantation suivante a été réalisée [10] :

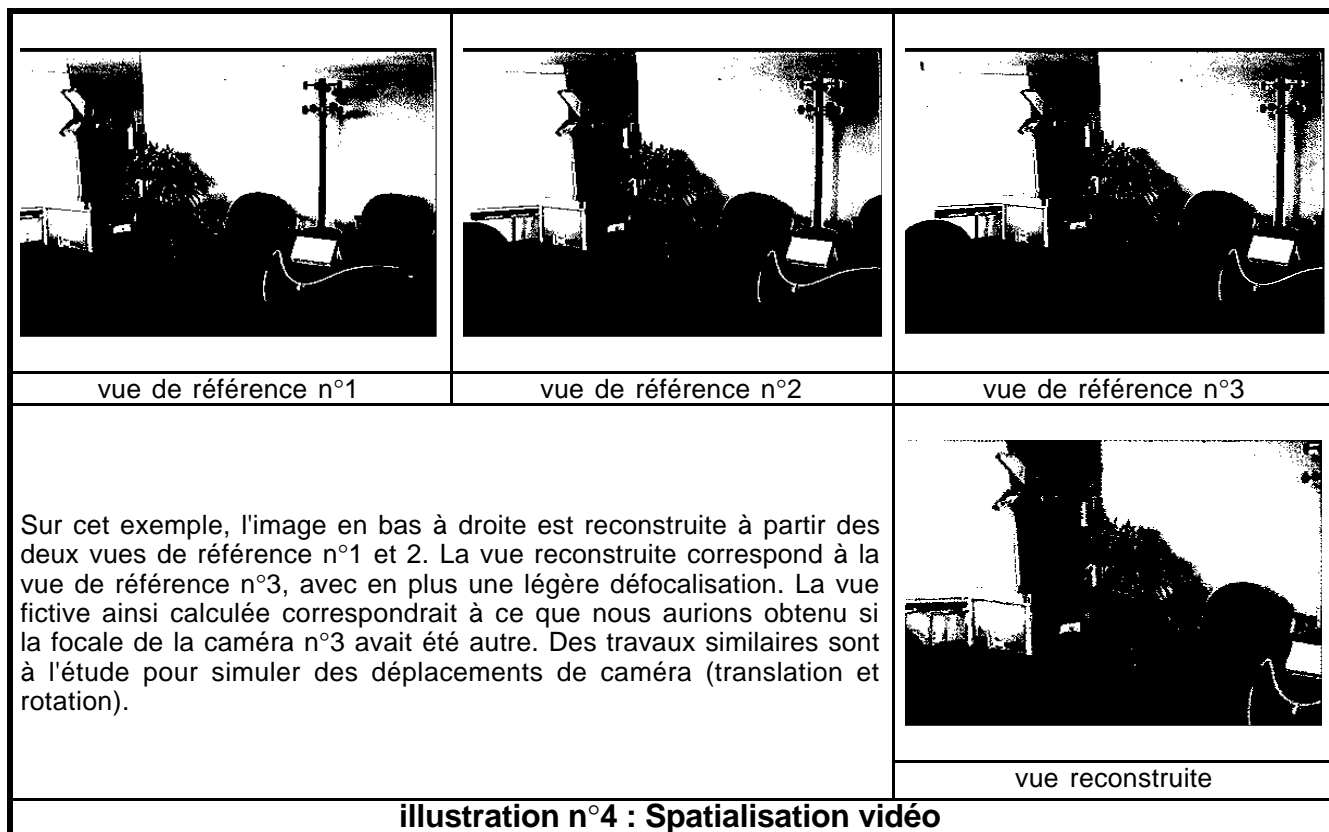
**ETAPE D'ANALYSE.** Mise en correspondance de triplets de points dits d'appui afin de calculer les paramètres de pseudo-calibration (quelques dizaines de points caractéristiques). Il est à noter que ces paramètres sont spécifiques à une vue, et sont de plus relatifs aux deux autres vues utilisées pour la reconstruction. Ces paramètres sont donc fonction de la configuration des trois caméras.

**ETAPE DE SYNTHÈSE.** Mise en correspondance dense des deux vues qui ne sont pas à synthétiser, puis reconstruction de la troisième vue à partir des paramètres dits de calibration (obtenus à l'analyse) et de la mise en correspondance réalisée précédemment. Afin de restituer une image aussi proche que possible de l'originale, un post-traitement est nécessaire car il existe des conflits dans certaines zones de l'image (i.e. certains points ont plusieurs valeurs de luminance, d'autres aucune,...).

A ce stade, cette approche permet de reconstruire une troisième vue d'une scène à partir de deux autres vues à la condition que ces trois vues existent, y compris la troisième (i.e. la vue à reconstruire). La connaissance de cette troisième vue est actuellement indispensable pour d'une part calculer les paramètres dits de calibration qui interviennent dans les relations trilineaires (étape d'analyse), et d'autre part valider l'algorithme. Cette limite est la contrepartie du non-passage par une phase explicite de calibration 3D. A ce niveau, bien que cette approche puisse apporter un gain en terme de compression (i.e. possibilité de transmettre une vue à l'aide de quelques paramètres seulement, dans la mesure où les vues de référence sont connues du récepteur), elle ne répond pas entièrement aux objectifs fixés ci-dessus : au-delà de l'aspect communication, être en mesure de reconstruire une vue éventuellement inexistante. Il est néanmoins possible d'agir à deux niveaux lors de l'étape de synthèse pour reconstruire de nouveaux points de vue:

- en modifiant les valeurs d'un ou plusieurs paramètres dits de calibration,
- en changeant une ou les deux vues de référence.

Ces deux possibilités permettent effectivement de reconstruire une vue fictive mais il est encore difficile de maîtriser a priori le point de l'espace qui sera ainsi atteint (illustration n°4).



**illustration n°4 : Spatialisation vidéo**

### 3.2 A partir de $n$ vues de référence

Afin de couvrir entièrement la salle de réunion avec une précision suffisante, il sera nécessaire d'extrapoler l'approche présentée précédemment afin qu'elle soit opérationnelle pour un ensemble de  $n$  vues ( $n \gg 3$ ). Il s'agira ensuite de calculer pour chaque vue de référence l'ensemble des paramètres de calibration précités (relatifs à un ou plusieurs couples de vues sélectionnés parmi les  $n-1$  autres vues). Les paramètres dits de calibration d'un point de vue fictif de la scène seront ensuite estimés, par interpolation, à partir des paramètres de calibration associés aux vues de référence. Une fois que ces paramètres, associés à la vue fictive souhaitée, auront été calculés, il sera alors possible de reconstruire un point de vue virtuel en choisissant également les deux vues de référence les plus adéquates parmi les  $n$ .

## 4. Conclusion et perspectives

*TRAVI* est un projet de télé-virtualité. Ce projet vise à mettre en place une structure de télécommunication audio-vidéo entre plusieurs sites distants via des liaisons bas-débit afin que différentes personnes puissent converser confortablement, en réduisant au maximum l'impression de distance grâce à des outils de la réalité virtuelle. Les traitements vidéo interviennent à plusieurs niveaux dans ce projet: le clonage et la spatialisation vidéo. Afin de garantir un niveau de réalisme visuel satisfaisant, les images virtuelles sont créées à partir d'une base d'images issues d'acquisitions réelles.

## Références

[1] Synthèse d'environnements sonores tridimensionnels pour la production audio ou multimédia et les interfaces homme-machine interactives.

J.-M Jot

L'Interface des mondes réels & virtuels, 21-24 Mai 1996, Montpellier, France.

[2] Televirtuality Project: Cloning and Real Time Animation System

Institut National de l'Audiovisuel

<http://www.ina.fr/INA/Recherche/TV/TV.en.html>

[3] Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at low bit-rates,

A. Eleftheriadis and A. Jacquin

Signal Processing: Image COMMUNICATION 7 (1995) 231-248.

[4] Télé-asservissement d'un modèle CYBERWARE

Rapports Internes EURECOM

Part. I: P. Abel, L. Ponce, Y. Rahmoun et N. Romanetti ( EURECOM), Déc. 95.

Part. II: L. Dere, C. Rami et R. Trento (ESSI), Mai 95.

[5] Active Contours for Lipreading: Combining Snakes with Templates

S. Horbelt and J.-L. Dugelay

Quizième colloque GRETSI - Juan-les-Pins - du 18 au 21 Septembre 1995, France.

[6] Simulation de chirurgie craniofaciale et réalité virtuelle

H. Delingette, G. Subsol, S. Cotin & J. Pignon

L'Interface des mondes réels & virtuels, pp. 399- 408, 7-11 février 1994, Montpellier, France.

[7] On the production and Perception of Audio-Visual Speech by Man and Machines

C. Benoit

Int. Symp. on Multimedia Communications & Video Coding, Polytechnic University, NYC, Oct. 11-13, 1995.

[8] Human Facial Motion Analysis and Synthesis with Application to Model-Based Coding

K. Aizawa, C.S. Choi, H. Harashima and T.S. Huang,

MOTION ANALYSIS AND IMAGE SEQUENCE PROCESSING, Chp. 11.

Edited by M. Ibrahim Sezan and Reginald L. Lagendijk

Kluwer Academic Publishers, 1993.

[9] Projective Structure from Uncalibrated Images: Structure from Motion and Recognition

A. Shashua

IEEE Trans. on PAMI, 1994.

[10] Spatialisation vidéo

K. Fintzel and J.-L. Dugelay

In Proc. CORESA'96 Conf. CORESA'96, Fév. 96, Cnet Grenoble, France.