ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0067

EDITE ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité " Traitement de l'Image et du Signal "

présentée et soutenue publiquement par

Marc DECOMBAS

le 22 11 2013

Compression vidéo bas débit par analyse du contenu

Low bitrate video compression by content characterization

Directeur de thèse : **Frédéric DUFAUX**
Co-encadrement de la thèse : **Béatrice PESQUET-POPESCU, Erwann RENAN, François CAPMAN**

Jury

M. Ferran MARQUES, Professeur, Dept. STC, UPC BarcelonaTech	Rapporteur
M. Dave BULL, Professeur, Dept. of Electrical & Electronic Eng., University of Bristol	Rapporteur
M. Janusz KONRAD, Professeur, Dept. of Electrical and Computer Eng, Boston University	Examinateur
M. Marc ANTONINI, Dir. recherche CNRS, Lab. I3S, Université Nice-Sophia Antipolis	Président de Jury
M. Frédéric DUFAUX, Dir. recherche CNRS, Dept. LTCI, TelecomParisTech	Directeur de thèse
Mme Beatrice PESQUET-POPESCU, Dir. recherche CNRS, Dept. LTCI, TelecomParisTech	Co-directrice de thèse
M. François CAPMAN, Ingénieur Docteur, Unité de recherche, TCS	Encadrant
M. Erwann RENAN, Ingénieur, Lab. MMP, TCS	Encadrant

T H È S E

Télécom ParisTech
école de l'Institut Mines Télécom – membre de ParisTech
46, rue Barrault – 75634 Paris Cedex 13 – Tél. + 33 (0)1 45 81 77 77 – www.telecom-paristech.fr

Remerciements

Je remercie tous les gens qui m'ont permis de faire carrière dans la recherche, un des rares endroits où l'on considère que « Le génie, c'est l'erreur dans le système » Paul Klee
Je tiens à remercier, ma mère pour m'avoir donné l'énergie et la force de poursuivre un tel projet à la manière Décombas. C'était cool, j'espère que tu es fier de moi !
Je remercie aussi, mon amie Nina, pour avoir fait l'énorme effort de comprendre ma thèse, de m'avoir tant de fois aidé à écrire des phrases françaises, pour avoir corrigé sur mon anglais, de m'avoir permis d'être plus rigoureux!
Mon directeur de thèse, Frédéric, pour avoir su me conseiller, m'encadrer et m'épanouir durant ces 3 ans de doctorat. Merci aussi pour le soutien que vous avez su m'apporter ! Merci encore.
Ma codirectrice de thèse, Béatrice, pour avoir proposé ce sujet très innovant et pour ses très nombreuses bonnes idées qui m'ont permis de toujours avancer. Merci bien.
Mon encadrant, Erwann, pour m'avoir donné envie de faire une thèse, fait le nécessaire pour que je la commence,pour m'avoir aidé à bien comprendre les enjeux de ma thèse et aidé à relativiser durant les moments difficiles. Merci beaucoup Erwann !
Mon encadrant, François, pour ses très bonnes idées, sa patience, son sang-froid et son soutien, jusqu'au dernier moment. Je ne l'oublierai pas, merci !
Mes rapporteurs, Dave Bull et Ferran Marques, et examinateurs, Janusz Konrad et Marc Antonini pour avoir lu et commenté avec attention mon manuscrit, mais aussi pour toutes ces questions très pertinentes durant ma soutenance. J'ai réellement apprécié vous avoir comme membre de mon jury. Merci encore !
J'espère que nous aurons l'occasion de collaborer à l'avenir.
Mon chef de laboratoire, Bruno, pour m'avoir permis de faire un stage dans le laboratoire, qui mena à un doctorat. Merci encore pour ton management et la liberté d'action que tu m'as laissés, chose que j'ai plus qu'appréciée et qui, je le sais, me manquera ! J'espère sincèrement que nous pourrons continuer à travailler ensemble et bien évidemment à jouer au volley !
Notre responsable des projets amont, Marc, pour m'avoir d'abord rencontré et écouté, pour toutes ces discussions très intéressantes et enrichissantes, pour tous ses conseils, d'être venu à ma soutenance et pour m'avoir donné les outils pour réussir dans ma vie professionnelle. J'espère sincèrement que nous nous reverrons !
Ma responsable des Ressources humaines, Francelise, pour tout le temps que vous avez consacré à mon cas, pour m'avoir réellement aidé à créer un plan de carrière cohérent à un moment où j'envisageais d'élever des chèvres dans la creuse à la suite de ma thèse !
Je remercie mes stagiaires, Simon, François L., François S., Younous, Elsa pour leur travail et leur aide
Je remercie les collègues de Thales, Bruno, Cyril, Benoit, Marc, Ben, François, Rachid, Bertrand, Seb et ceux de télécom, Jean Hugues, Jean Robby, Jean Sylvain, Eric Le Rouge, Paolo, Emilie, BP, Mounira pour m'avoir donné un cadre de travail plus que chaleureux. Merci à ma partenaire de danse Elina pour les lundis soir plus que sympathique !
Ainsi que les autres copains pour m'avoir suivi quand plus personne ne voulait me suivre (John, Boris, Grosclem, Kinder, Pusum, Ju, Rémy, Elo, Manue, Soraya)
Merci aussi à Aurel pour le pot de thèse !
Merci à ceux que j'ai oublié de citer (qui ne vont pas me remercier) !

Et finalement, comme le dirait Albert
« L'imagination est plus importante que la connaissance » Albert Einstein

1. Introduction, enjeux et contributions

1.1. Le contexte

1.1.1. Introduction et application professionnelle

Les pays et les entreprises ont depuis quelques années décidé de faire évoluer leur façon de protéger les citoyens. Les solutions de vidéosurveillance ont émergé un peu partout dans le monde. Les pays développés comme les États-Unis ou l'Angleterre ont une politique très poussée sur ce type de protection. Les pays en voie de développement considèrent également la vidéoprotection comme une solution pour faire baisser la criminalité : en exemple, Mexico était la ville la plus dangereuse du pays et a mis en place un projet sur 3 ans de vidéosurveillance pour un total de 460 millions de dollars. Ce projet entre Thales, Mexico et Telmex, avait pour but de permettre à la ville de gérer un grand nombre de risques, tels que la délinquance, le terrorisme, les attaques de sites stratégiques ou encore les risques naturels. Cela a nécessité le déploiement de 8080 caméras et capteurs.

Les grandes entreprises voient aussi la solution de vidéoprotection, comme une bonne manière de protéger les biens et leurs employés. Celle-ci leur permet de répondre plus rapidement à un vol, une attaque, ou encore un incendie. Des groupes comme Thales développent des solutions de sécurité civile et urbaine depuis 15 ans pour les États, qui ont pour but d'augmenter l'efficacité des agents sur le terrain et de diminuer le coût de surveillance : un exemple simple peut être représenté par la surveillance des métros. La ville de Paris possède 14 lignes avec 300 stations réparties sur un total de 213 km de rames [RATP]. La protection des 1.4 milliard de passagers contre les agressions, les voleurs à la tire et la fraude n'est humainement pas possible. La RATP a donc développé les solutions de vidéosurveillance en déployant 8 200 caméras sur son réseau et 18 000 dans les autobus [Parisien 2012] pour tenter de résoudre ce genre de problème.

Les solutions de vidéoprotection permettent aussi d'accéder à des zones trop dangereuses pour l'être humain. On peut citer la protection de camps militaires à l'aide de capteurs déposés, la récupération d'informations avec des drones volants ou terrestres, etc. Toutes ces solutions ont pour but d'être plus réactives et de mieux coordonner les actions en cas de crises telles qu'une attaque terroriste ou une catastrophe naturelle. Cette demande croissante a permis la réduction des coûts de capteurs et la démocratisation de la caméra dans notre environnement.

La Grande-Bretagne est maintenant dotée de plus de 1.85 million de caméras [Reeve 2011], [Dailymail 2011], et les États-Unis, 30 millions [Popularmechanics 2009], [Vlahos 2008]. La France a lancé sur Paris un grand plan de vidéoprotection en 2010 [Figaro 2010] et, en 2013, autorise la vidéoverbalisation [Lepoint 2013].

Des solutions plus individuelles se développent avec par exemple la société iwatchlife [Iwatchlife 2013]. Cette entreprise propose une solution de vidéoprotection contre les effractions et autres évènements anormaux et vous les transmet sur votre mobile. D'autres sociétés, telles que SecuriteEnEntreprise, proposent de la protection temporaire de sites pour des évènements comme des festivals, des livraisons ou encore des chantiers, dans lesquels des solutions de vidéoprotection [Securiteentrepise] peuvent être très utiles.

Cependant, la quasi-totalité des solutions proposées utilise des encodeurs vidéo traditionnels qui ont été développés pour le grand public. Principalement utilisés pour la transmission vidéo ou

l'enregistrement de films, ces outils cherchent à maintenir la qualité globale de l'image et à obtenir un rendu agréable. La qualité de l'image n'est pas indispensable pour les applications professionnelles, dont les priorités sont plus de détecter des évènements, de comprendre ce qui se passe et de fonctionner dans toutes les conditions.

1.1.2. Les conditions

1.1.2.1. Capteurs nombreux et/ou de bonne qualité

Les caméras IP ont vu leurs coûts fortement diminuer avec la demande croissante [Mantratec]. La densification et l'augmentation des capacités réseau ont permis de connecter toutes ces caméras en réseaux. On rappelle qu'en réseau filaire, les capacités de transmission sont de 10 Gb/s [Landa 2006] [Carte_Reseau] et parallèlement différents types de réseaux non filaires se sont développés. On peut évoquer le Bluetooth qui permet une communication de courte distance entre deux objets, le WIFI qui permet la connexion d'un ordinateur ou d'un portable à un réseau public ou privé ou encore les solutions mobiles (edge, 3G, 4G). Toutes ces techniques permettent l'interconnexion des objets avec des capacités de transmission toujours plus importantes.

Cette capacité des réseaux, la définition des écrans et des capteurs ainsi que la volonté d'avoir aussi des images d'une qualité toujours meilleure a permis de passer de formats d'image 640x480 Video Graphic Array (VGA) en 1987 [VGA] à des formats tel que le full-HD (1920 x 1080) ou encore prochainement le 4K(3840 × 2160).

Ces vidéos exigeant toujours plus de débit peuvent être transmises dans des lieux ayant un réseau bien développé. Cependant, le débit disponible peut être très variable en fonction des lieux. Le robustification des capteurs et l'augmentation des capacités des batteries ont permis de retrouver des caméras haute définition sur des véhicules, des personnes, des drones ou encore directement déposées à même le sol.

1.1.2.2. Très faible débit disponible ou fort coûteux

On a vu précédemment que les capacités des réseaux étaient de plus en plus importantes, mais pas nécessairement bien réparties. Les solutions de réseau téléphonique de 4e génération (4G), de réseau individuel (WIFI) et de fibres optiques sont disponibles dans les grandes villes, mais très rapidement, on peut se trouver des débits de plus faible capacité dans les campagnes des pays développés. Pour les pays en voie de développement, les solutions filaires étant très coûteuses à mettre en place, les solutions non filaires ont été favorisées. Les solutions GSM sont donc présentes à peu près partout dans le monde, à condition qu'il y ait une certaine densité humaine. Cependant, dans des zones à très faible densité de population, telles que les déserts, les grandes forêts, les grands lacs et les mers, voire les océans, il n'existe pas de moyen de transmettre des informations.

Les solutions présentées sont proposées par des opérateurs publics et sont accessibles à tous. Or, pour les applications professionnelles, il est parfois préférable d'avoir son propre réseau afin de contrôler la transmission de l'information d'un bout à l'autre de la chaîne et ainsi sécuriser l'information. De plus, les solutions professionnelles doivent être fonctionnelles n'importe où. L'utilisation des réseaux publics est donc envisageable, à condition d'être utilisables en toute situation et dans le cas où l'information transmise n'a pas besoin d'être fortement sécurisée. Elles utilisent donc de la transmission par radio fonctionnant avec un faible débit ou encore satellitaire, dont le coût est très élevé.

1.1.2.3. Zones peu accessibles ou dangereuses

Les professionnelles doivent être capables d'agir en tout lieu et en toute situation.

En plus d'avoir peu de débit, certaines zones sont difficiles d'accès. Prenons l'exemple de la haute montagne où le débit est faible et les lieux sont difficiles d'accès. L'Institut de Formation et de Recherche en Médecine de Montagne (IFREMMONT) [Iffremont] cherche à développer des solutions d'e-médecines afin de permettre l'assistance médicale à distance. On peut donc envisager une transmission vidéo permettant à un médecin d'analyser la situation et de pouvoir donner les conseils nécessaires en attendant l'arrivée des secours.

Les domaines maritimes aériens et spatiaux sont également des zones peu accessibles et très vastes, donc difficilement raccordables. On imagine difficilement déployer une protection humaine pour surveiller un désert ou les frontières d'un pays.

Enfin, certaines zones peuvent aussi être dangereuses. On peut citer deux cas, les zones de guerre et les zones ayant subi une catastrophe naturelle ou humaine. Dans ces deux derniers cas, on peut envisager de déployer une ou des équipes de personnes, mais celles-ci seront constamment en danger soit d'être attaquées en cas de guerre, soit de subir l'environnement dans le cas de catastrophe naturelle. Or, dans tous les cas, les solutions de vidéoprotection permettent de diminuer les risques, mais pour cela, il faut réussir à transmettre l'information nécessaire à la prise de décision.

1.1.2.4. Besoin de récupérer l'information pour prise de décision

Avec l'augmentation des capteurs, le besoin de fonctionner en toute situation et en tout lieu, et la difficulté de transmettre l'information, il est nécessaire de définir ce qu'il est important de transmettre et à quoi va servir cette information.

L'information est transmise à un centre de prise de décisions. Elle doit donc permettre d'analyser le problème afin de comprendre ce qu'il se passe et les enjeux en cours. Après analyse, le centre de contrôle doit pouvoir prendre des décisions, donner des ordres et des conseils afin de réussir à modifier et à améliorer la situation sur le terrain. Les prises de décisions au centre de contrôle permettent aussi de réduire le coût d'une opération en évitant, par exemple, le déplacement d'un hélicoptère pour secourir quelqu'un en haute montagne ou d'envoyer une deuxième fois un drone en opération de récupération d'information au-dessus d'un territoire ennemi pour compléter la première mission. Cela permet également de diminuer le risque humain en évitant de déplacer une équipe dans un territoire dangereux ou en coordonnant mieux une équipe au sol. À partir de ces considérations, quelques scénarios d'applications vont être présentés en détail.

1.1.3. Les scénarios

1.1.3.1. Capteurs déposés

Le premier scénario présenté est celui des capteurs déposés.

Ce type de capteur permet de remonter une alerte dans une zone non surveillée par un être humain. Ces zones peuvent être dangereuses, suspectes ou très étendues.

Dans ce type de scénario, en terme de matériel, il est usuel d'utiliser des capteurs vidéos combinés à d'autres capteurs comme des capteurs sismiques, lasers et audios.

La raison est double : les capteurs doivent pouvoir remonter une alerte et ils seront plus robustes s'ils utilisent différents stimuli. La deuxième raison est qu'un capteur vidéo consomme quantité d'énergie et, comme il peut être abandonné pendant plusieurs mois, il doit posséder un système de gestion énergétique très optimisé. Le plus simple est donc de laisser le capteur vidéo endormi et de le réveiller lorsqu'il y a un stimulus sur les autres capteurs. Outre la contrainte d'énergie, il y a aussi une contrainte humaine. Le temps de mise en route et de dépose doit être rapide, car le capteur est, par définition, déposé dans une zone où l'on ne peut pas laisser longtemps un être humain. Il faut qu'il soit également robuste relativement aux différents scénarios et conditions ainsi que, rapide à paramétrer.

Voici un exemple de scénario : une zone est suspectée par la police d'être le lieu d'activités illégales. Elle se trouve dans une région peu couverte en termes de réseau, et il n'est pas envisageable de laisser une équipe pour des raisons de discrétion aussi bien que d''économie. On peut installer une caméra avec batterie, mais sans possibilité de remonter toute l'information. Il devient donc difficile de savoir quand venir récupérer le matériel sans prendre le risque d'être découvert. D'autre part, une transmission permanente rend le dispositif plus facilement détectable et consommera beaucoup d'énergie. Une solution de transmission vidéo bas débit adaptée au contenu peut résoudre ce problème.

1.1.3.2. Drones

Dans ce deuxième scénario, nous nous intéressons aux drones.

Les drones sont des véhicules sans pilote humain, aériens ou terrestres, et utilisés pour la surveillance et la cartographie de zones dangereuses. Les capteurs vidéo utilisés sont souvent de très haute définition afin de pouvoir récupérer au mieux le maximum d'informations nécessaires à l'analyse. Les drones partent en reconnaissance dans des zones suspectes avec un réseau pouvant avoir un débit très faible. Il est cependant possible d'imaginer que le drone revienne dans une zone mieux couverte pour envoyer une partie des informations.

Par exemple dans un cas normal, un drone suit son plan de vol, et l'on vérifie que les données acquises sont utiles une fois le drone rentré à la base. Or, si l'on remarque après analyse qu'il manque des informations pertinentes, un nouveau plan de vol doit être fait, ce qui est coûteux et risqué puisque le drone a pu être repéré par l'ennemi survolé ou a traversé une zone turbulente avec le risque d'être détruit. Avec une solution de transmission adaptée au contenu et à bas débit, le drone peut envoyer une séquence de la zone d'intérêt, ce qui permettra au centre de commandement d'analyser cet échantillon de résultats et de corriger le plan de vol si nécessaire.

1.1.3.3. Caméra embarquée véhicule ou fantassin

Le dernier scénario présenté est celui des caméras embarquées sur un véhicule ou un être humain.

Assez proche du cas du drone dans le concept, il reste très différent dans les applications. On peut utiliser des caméras embarquées afin que le porteur de caméras embarquées puisse mieux détecter les activités dans son environnement, voire transmettre ces informations à un centre de commandement, qui pourra les visualiser et l'aider dans la prise de décision. Les caméras embarquées permettent aussi d'analyser a posteriori des interventions afin de mieux comprendre ce qui s'est passé et de rectifier certaines erreurs pour améliorer les prochaines missions.

Une transmission en temps réel ou légèrement différé permet d'améliorer ces actions durant l'intervention. Les professionnels (policiers, pompiers ou fantassins) sont ainsi équipés de caméras dans leurs interventions. Le matériel utilisé peut être composé d'une multitude de capteurs de haute résolution (caméra visuelle, caméra infrarouge ou caméra thermique), ce qui peut donner une quantité importante d'information. Le réseau ne pouvant pas nécessairement la gérer, cela peut devenir une contrainte dans le cas de la transmission au centre de contrôle. Une autre contrainte est qu'il y a souvent plusieurs personnes trop proches physiquement qui utilise les mêmes antennes de transmission et sature le réseau. De plus, comme il est souligné précédemment, au cours des applications professionnelles, il est préférable de maîtriser dans la mesure du possible toute la chaîne de transmission afin de ne pas être dépendant d'un opérateur public, mais aussi d'être efficace en toute situation. Cela fait que les outils de transmission sont embarqués et ne sont pas nécessairement d'une grande capacité de transmission. Prenons en exemple, une équipe de pompiers qui part en intervention dans un immeuble en flammes. Chaque pompier est équipé d'une caméra thermique et d'une caméra visuelle. Le centre de commandement déplace son équipe en écoutant et en analysant les conversations radio. Une transmission vidéo adaptée au contenu permettrait au centre de commandement de visualiser automatiquement l'action et les problèmes des personnes sur le terrain dans les zones sensibles et ainsi de mieux diriger, d'organiser et de sécuriser l'équipe en intervention.

Dans ces trois scénarios, il n'est pas nécessaire d'accéder à l'ensemble de l'information, mais l'on constate qu'une partie seulement est utile et permet de prendre des décisions.

Actuellement, la solution disponible sur le marché la mieux représentée est l'encodeur grand public H.264/AVC.

1.2. L'encodage

1.2.1. Les motivations de l'encodage dans le cas des applications grand publics

Afin de mesurer les motivations de l'encodage, prenons l'exemple d'un film de 1 h 30 en full-HD en 25 images par seconde. Le poids du fichier va être calculé ainsi : Nb_Images_par_Sec x Nombre_ de_Sec x Nombre_de_Canaux x Dim_Spatial x Dynamique = 25 x 90 x 60 x 3 x 1920 x 1080 x 1 = 838,808 Go. On voit donc que sans encodage, un simple film est presque impossible à stocker et à diffuser. Pour des raisons commerciales et techniques, les films doivent avoir un poids inférieur au support physique existant. Le CD a une capacité de 700 Mo, le DVD de 4,37 Go à 8,54 Go et le Blu-ray de 7 à 100 Go. On voit que, même avec les meilleures méthodes de stockage physique, il faudrait 8 Blu-ray pour un film de 1 h 30, ce qui impliquerait de changer de disque environ toutes les 12 minutes.

Lors de l'encodage, l'objectif est de maintenir au mieux la qualité globale du film et d'avoir la meilleure qualité possible en fonction du support ou de la capacité de transmission. Afin de mesurer la qualité globale d'une image, des métriques objectives telles que le Pick Signal to Noise Ratio (PSNR) ou encore le Structural SIMilarity (SSIM) ont été établis. Les méthodes d'encodage cherchent à maximiser la qualité en fonction du débit disponible.

1.2.2. Les méthodes de réduction de l'information

Les séquences vidéo ont une très forte quantité d'informations redondantes, aussi bien dans le domaine spatial que dans le domaine temporel. Les techniques de compression se basent sur la forte corrélation entre les pixels, aussi bien spatiale, où les pixels adjacents sont similaires, que temporelle, où les pixels des images passées et futures sont également très proches.

Les méthodes d'encodage cherchent à utiliser une combinaison d'encodage intra et inter. Le codage intra utilisera la corrélation spatiale entre les pixels, et le codage inter, la corrélation temporelle.

L'encodage inter est dépendant des images précédentes ou suivantes. En cas de perte d'une image pendant la transmission, il faudra attendre la prochaine image encodée en intra. On définit la notion de Group Of Pictures, comme un ensemble de trames regroupées répété périodiquement jusqu'à la fin de la vidéo encodée. Un GOP définit l'ordre dans lequel sont disposées les images à encodage (intra) et à encodage prédictif (inter).

1.2.2.1. Sous échantillonnage et interpolation

Une méthode simple pour réduire la quantité d'information consiste à faire un sous-échantillonnage de la source qui réduit les dimensions spatiales de la vidéo, et qui diminue ainsi le nombre de pixels à encoder. Il est aussi possible de sous-échantillonner l'aspect temporel en diminuant le nombre d'images par seconde. Afin d'avoir un résultat final aux dimensions spatiales et à la fréquence identique à la source, une étape d'interpolation est nécessaire pour permettre de synthétiser les parties supprimées.

L'œil humain étant plus sensible aux variations de luminosité qu'à celles de couleurs, la majorité des encodeurs utilise l'espace YUV, où Y est une composante de luminance et UV sont deux composantes de chrominance. Les composantes de chrominance sont sous-échantillonnées de différentes manières d'après le tableau suivant :

Appellation (YUV)	Luminance (Y)	Chrominance (U et V)
4 :0 :0	Pleine résolution verticale et horizontale	Pas de chrominance
4 :1 :1	Pleine résolution verticale et horizontale	¼ de résolution horizontale et pleine résolution verticale
4 :2 :0	Pleine résolution verticale et horizontale	½ de résolution horizontale et ½ de résolution verticale
4 :2 :2	Pleine résolution verticale et horizontale	½ de résolution horizontale et pleine résolution verticale
4 :4 :4	Pleine résolution verticale et horizontale	Pleine resolution horizontale et verticale

Tableau 1. Table de sous échantillonnage YUV

Deux représentations sont particulièrement utilisées en encodage, YUV 4 :2 :0 et YUV 4 :4 :4. Dans la première, la luminance est en pleine résolution, les chrominances sont sous échantillonnés par deux

verticalement et horizontalement, ce qui donne un quart de résolution et qui diminue fortement la quantité d'information à transmettre. Dans la deuxième représentation, la luminance et la chrominance sont en pleine résolution, ce qui permet d'avoir une source sans perte d'échantillonnage.

1.2.2.2. L'encodage intra

L'encodage intra basé sur la corrélation spatiale introduit la notion de trames I et utilise le fait qu'un pixel ou un groupe de pixels peut être prédit à partir de ses voisins. La trame I est encodée sans aucune référence sur les trames futures et passées. L'image est découpée en bloc de 8x8 pixels et, pour chaque bloc, une Transformation en Cosinus Discret (DCT) est appliquée. Cette transformation permet de passer du domaine spatiotemporel au domaine fréquentiel, où l'on pourra appliquer une quantification. Durant la quantification, on pourra supprimer les hautes fréquences, car celles-ci sont peu perceptibles par l'œil humain et peuvent représenter une importante quantité d'information. De manière générale, la quantification va permettre de réduire la quantité d'information en passant d'un ensemble de valeurs possibles à un autre plus faible.

En résumé, les images Intra sont donc des images fixes et indépendantes des autres types d'images. Chaque Groupe of Pictures (GOP) commence par ce type d'image.

1.2.2.3. L'encodage inter

L'encodage inter introduit les trames P et B et utilise le fait que, dans un voisinage temporel proche, les pixels ne changent pas énormément. La prédiction du mouvement est un puissant moyen de réduire la redondance temporelle entre les images et se fait à partir d'images de référence. Le concept est basé sur l'estimation du mouvement entre les images. Si tous les éléments de la scène ont des mouvements relativement simples, le mouvement entre une image et une autre peut être décrit avec un nombre limité de paramètres. Ces paramètres, appelés vecteurs de mouvement, sont encodés et transmis.

Les images P (Prédictive) contiennent des informations de différence résultant de la prédiction compensée de mouvement avec l'image I ou l'image P précédente. Elles servent aussi d'images de référence.

L'image B (Image à codage prédictif Bidirectionnel) contient des informations de différence avec les images I ou P passées et futures à l'intérieur du GOP. Afin de limiter la propagation d'erreurs de prédiction, les images B ne sont généralement pas utilisées en tant qu'images de référence.

La Figure 1 illustre les images I, P, B avec leurs prédictions ainsi que la notion de GOP. L'image I_1 ouvre le GOP et sert de référence pour l'image P_1 qui sert elle-même de référence pour P_2. L'image B_1 est prédite par l'image du passé I_1 et l'image du futur P_1. Pour B_2, P_1 sert de référence passée et P2 de référence future. I_1, B1, P_1, B_2, P_2 forment un GOP. I_2 initie le GOP suivant.

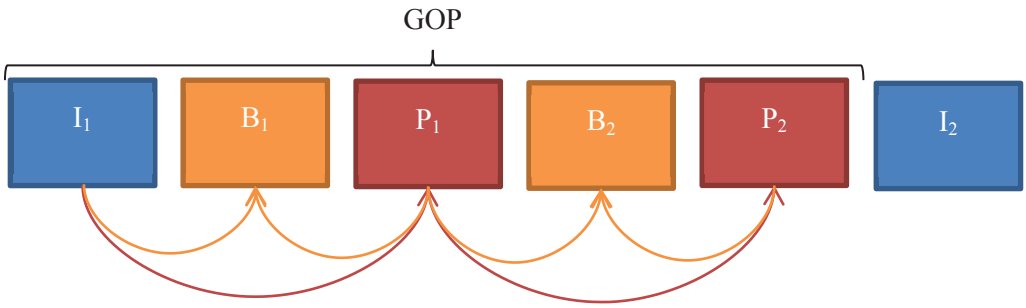


Figure 1 Images I, P, B avec leurs prédictions et notion de GOP

1.2.2.4. Le codage à longueur variable

Une fois les prédictions faites, il faut représenter les symboles à transmettre sur un certain nombre de bits. La méthode la plus simple est d'utiliser un codage à longueur fixe, où l'on attribue le même nombre de bits aux symboles, quels qu'ils soient. Ce nombre de bits est défini en fonction de la dynamique des symboles. À titre d'exemple, un pixel peut avoir une valeur entre 0 et 255 sur chacune de ses composantes RGB. Il faudra donc $2^8 = 256$ valeurs pour représenter les symboles, soit 8 bits par composante.

Le code à longueur variable est un code qui associe les symboles de la source à un nombre variable de bits et qui permet à la source d'être compressée et décompressée avec une erreur nulle. Ce sont des encodages sans perte. Le concept est d'associer un faible nombre de bits aux symboles les plus fréquents et, réciproquement, un nombre élevé de bits aux symboles moins fréquents. Les méthodes de codage utilisant ce principe sont le codage de Huffman [Huffman 1952], le codage Lempel-Ziv [Welch 1984] et le codage arithmétique [Witten 1987].

1.2.3. Les motivations de l'encodage pour les applications professionnelles et conclusion

Dans les applications professionnelles, l'encodage a pour objectif d'assurer le lien entre la source et le destinataire, et ce, quelles que soient les situations. Une perte de liaison peut entraîner de graves conséquences humaines (impossibilité de gérer une équipe sur le sol en terrain dangereux) et économiques (impossibilité de modifier le plan de vol d'un drone). Cela diffère des applications grand public où la perte d'information sera gênante, mais sans conséquence vitale. Dans les applications professionnelles, l'information qui sera transmise n'a pas pour but d'être de bonne qualité comme dans les applications grand public, mais de maintenir la sémantique à des fins d'analyse et de prise de décision.

On remarque que les encodeurs traditionnels ont été conçus pour répondre aux besoins du grand public, ce qui signifie d'avoir une image de meilleure qualité sur des supports pourvus de capacités toujours plus importantes.

Les encodeurs traditionnels ne sont pas directement adaptés aux applications professionnelles et peuvent mal fonctionner dans les cas où le débit est trop faible. En effet, si l'on souhaite maintenir une bonne qualité globale à très bas débit, l'ensemble de l'image sera dégradé et la sémantique de la scène perdue. Il n'y a pas de volonté de s'adapter au contenu et de faire de l'encodage sémantique.

1.3. Les objectifs

Dans un contexte de très bas débit, toute l'information est difficile à préserver. Il faut donc définir ce que l'on souhaite maintenir et trouver un moyen pour réduire intelligemment la quantité d'information. Il faudra donc déterminer un certain nombre de buts essentiels en fonction des contraintes et scénarios définis.

1.3.1. Rester compatible par rapport aux standards existants

Puisque la solution que nous souhaitons proposer doit pouvoir s'intégrer dans de nombreux produits déjà existants, il faut modifier au minimum les solutions existantes pour des raisons économiques et de compatibilités. Nous opterons donc pour un processus qui s'applique en amont et en aval de l'encodage traditionnel.

1.3.2. Définir ce qui est nécessaire pour interpréter une vidéo

Puisque le critère primordial de l'encodeur que nous souhaitons réaliser est le maintien de la sémantique et de l'interprétabilité de la vidéo, il faut savoir détecter les zones d'intérêt. Une fois ces zones détectées, il faudra allouer du débit et avoir une bonne qualité de rendue sur les objets d'intérêt afin de pouvoir les identifier et les caractériser. Il est aussi important pour des raisons de compréhension et d'interprétation de clairement identifier la position des objets et leurs mouvements relatifs tout en gardant assez d'information sur le fond pour pouvoir définir le contexte et l'environnement. L'utilisateur devra toujours rester critique vis-à-vis des traitements appliqués en les laissant par exemple facilement identifiables. Dans le cas de mauvaise détection d'un objet d'intérêt, cette précaution permettra à l'utilisateur d'appréhender l'erreur de l'encodeur et prendre ainsi les dispositions nécessaires pour la corriger en demandant à la source de renvoyer l'image en entier, sans traitement par exemple.

1.3.3. L'information transmise doit être modulable et fonctionner à faible débit

Une fois les zones d'intérêt détectées, il faut définir un moyen pour les séparer automatiquement du fond.

Une fois cette opération faite, il faut trouver la façon de concentrer au mieux l'information intéressante et supprimer l'information de moindre intérêt, qui est généralement le fond. Rappelons aussi qu'une partie de l'information du fond devra être maintenue afin de connaître le contexte. La diminution d'information non pertinente permettra d'avoir plus de flexibilité sur ce que l'on souhaite transmettre, à savoir une meilleure qualité sur les objets d'intérêt, une autre séquence, plus d'information sur le contexte. L'information à transmettre doit rester facilement modulable afin que l'utilisateur puisse facilement distribuer le débit disponible en fonction de ses besoins. Dans le cas où il faudrait transmettre une information supplémentaire, il est nécessaire de la représenter de la manière la plus compacte possible afin de limiter au maximum le surcoût.

1.3.4. Evaluable

Il faudra dans la mesure du possible pouvoir mesurer l'interprétabilité de la séquence. On pourra s'affranchir du fond et mesurer de manière séparée les déformations géométriques qui peuvent avoir lieu sur les objets et le maintien de leur position. Mesurer la qualité des objets d'intérêt reste essentielle afin d'estimer s'ils présentent encore assez de détails interprétables et s'ils sont encore caractérisables.

1.3.5. L'environnement

Connaissant les scénarios d'application, il faudra garder en tête que l'encodeur que nous proposons doit être facile à mettre en fonctionnement. Le paramétrage doit y être faible, voire adaptatif. Le codeur doit être capable de fonctionner dans tous les contextes, car il n'y a pas d'information connue à priori sur l'environnement et sur les scénarios. Une définition anticipée ou imaginée préalablement du scénario, du contexte ou de l'environnement peut être dangereuse, car il y a toujours le risque de se trouver dans une situation inédite et imprévisible.

1.3.6. Compression très bas débit par zone d'intérêt

N'ayant pas de scénarios prédéfinis, ni de matériels envisagés, nous nous sommes fixé l'objectif théorique de transmettre de la vidéo dans une fenêtre de 300 à 600 kbit/sec.

Aucune contrainte n'a été faite sur la complexité, la consommation d'énergie, la définition de la caméra ou encore le type de capteur (infrarouge, visible, etc.). Nous proposons donc une méthode de compression bas débit par zone d'intérêt répondant aux différentes problématiques et dans les conditions liées aux différents scénarios. Cela nous permettra de faire une étude de faisabilité.

1.4. Compression vidéo par redimensionnement

1.4.1. Le principe

Nous proposons ici de faire de la compression vidéo par redimensionnement.

L'idée est de supprimer une partie des informations dans la vidéo afin de réduire le débit.

À partir de la vidéo originale, une réduction spatiale est effectuée sur celle-ci. La vidéo réduite est alors encodée à l'aide d'un encodeur traditionnel et les informations permettant de revenir aux dimensions spatiales d'origine sont également encodées. En ce qui a trait au décodeur, la vidéo réduite et les informations de redimensionnement spatial sont décodées et combinées afin de revenir à une vidéo aux dimensions d'origine. Les zones supprimées pourront être synthétisées si cela aide à la compréhension de la scène. La Figure 3 résume le principe de cet encodage.

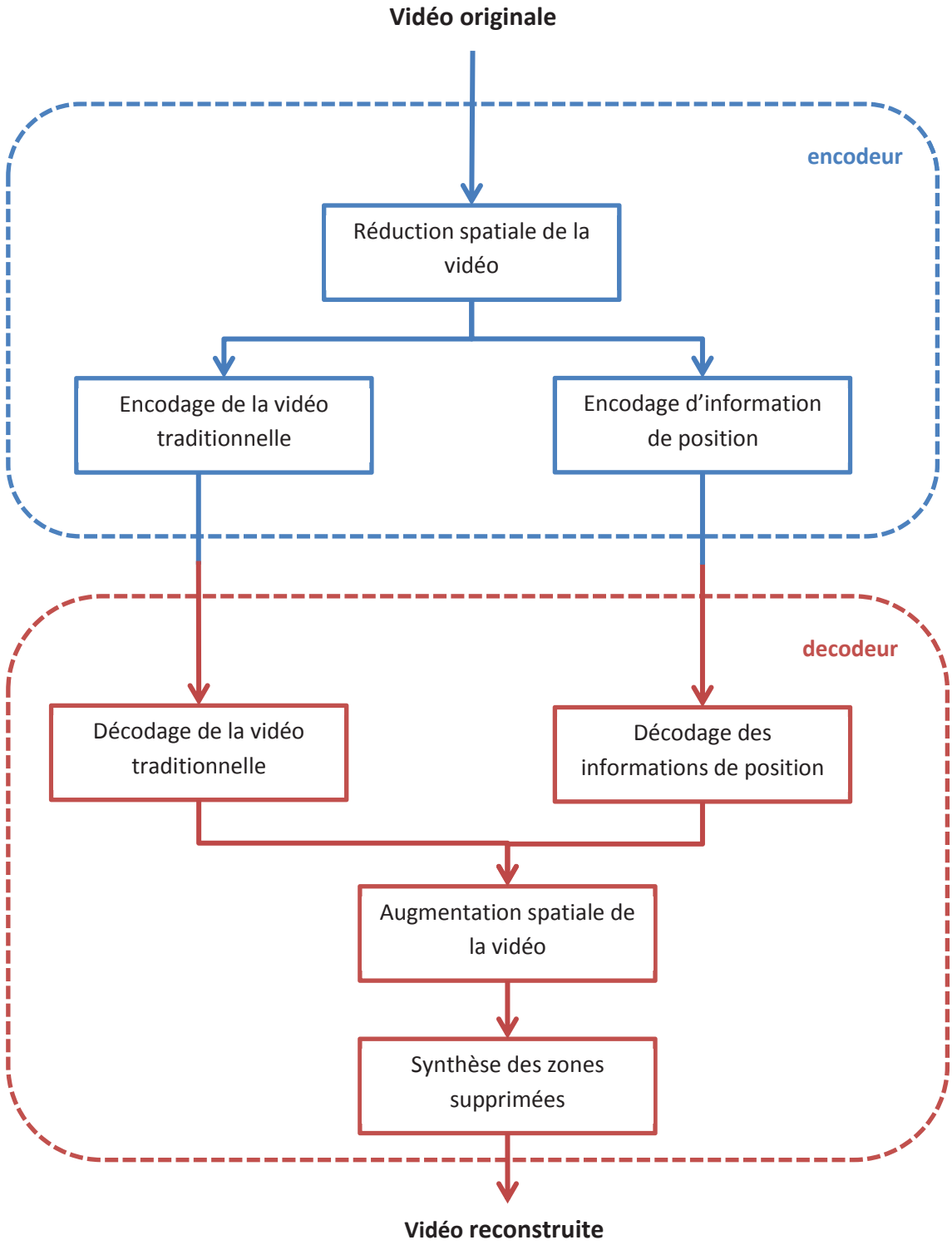


Figure 2 Schéma de compression vidéo par redimensionnement

1.4.2. Les briques élémentaires

Afin de parvenir à la réalisation de cet encodeur, plusieurs briques algorithmiques élémentaires sont nécessaires.

1.4.2.1. Un outil pour détecter les zones d'intérêt

Un outil de détection de zones d'intérêt est nécessaire. Deux grands types d'approche existent, l'extraction de fonds et les cartes de saillance.

Les outils d'extraction de fond cherchent à modéliser le fond et définissent les objets d'intérêt comme étant ce qui n'est pas le fond. Ces approches fonctionnent bien dans le cas de caméra fixe et permettent la plupart du temps d'obtenir un détourage précis des objets d'intérêt. Cependant, le résultat est binaire (est le fond ou ne l'est pas). Afin de savoir ce qui est le fond, il est souvent nécessaire d'avoir une phase d'apprentissage. Ces approches peuvent être sensibles au changement de luminosité et ne peuvent pas fonctionner si le fond change trop rapidement. Les cartes de saillance cherchent à modéliser le regard humain afin de prédire où un utilisateur va regarder. Elles fonctionnent dans de nombreux cas et nécessitent peu de conditions initiales. En résultat, elles donnent une probabilité pour chaque pixel d'être regardé. L'inconvénient de ces méthodes est que le résultat est souvent flou et le contour de la zone d'intérêt n'est pas lié au contour de l'objet d'intérêt.

On choisit donc les cartes de saillances étant donné les contextes d'application, car celles-ci ne nécessitent pas d'apprentissage définissant automatiquement ce qui est intéressant et peuvent donc fonctionner dans de nombreux cas.

1.4.2.2. Une méthode de redimensionnement spatiale

Du moment que l'on peut connaît les zones d'intérêt, l'information peu pertinente est supprimée en utilisant une méthode de redimensionnement spatial. Il en existe au moins cinq : le « resizing », le « cropping », le « warping », le « data pruning »et le « seam carving ».

Le « resizing » est une méthode qui réduit uniformément une vidéo en utilisant une approche pyramidale. À chaque pixel réduit à l'échelle « n » est associée la valeur moyenne d'un ensemble de pixels à l'échelle « n-1 ». L'intérêt de cette méthode est qu'elle est très simple à appliquer. Cependant, elle ne s'adapte pas au contenu et réduit aussi les objets d'intérêt. On perd donc des détails sur ceux-ci. Même si des informations supplémentaires sur les objets d'intérêt sont transmises, cette approche est difficilement réversible.

Le « cropping » est une méthode qui cherche à supprimer de l'information sur un bord de l'image. Elle est aussi très simple à appliquer et la quantité d'information à transmettre pour revenir aux dimensions d'origine est très faible. Cependant, cette approche ne fonctionne bien que si les objets d'intérêt ne sont pas sur les bords, et elle ne s'adapte pas au contenu.

Le « warping » se base sur un quadrillage de l'image, et chaque quadrant est redimensionné en fonction de sa saillance. Cette approche de redimensionnement à l'avantage d'être liée au contenu et de moins réduire les zones intéressantes. Cependant, la déformation peut être fortement calculatoire et difficilement modélisable.

Le « data pruning » consiste à supprimer des lignes droites dans l'image en fonction de leurs saillances. Cette approche a le mérite d'être facile à calculer, de s'adapter au contenu et de permettre facilement d'encoder l'information de redimensionnement. Cependant, elle révèle ses limites rapidement, à savoir que si les zones d'intérêt ont des formes complexes, car les lignes droites ne pourront pas éviter les objets d'intérêt.

Pour finir, la dernière méthode de redimensionnement présentée est le « seam carving ». Cette approche supprime des lignes courbes (seams) dans l'image en fonction de leur saillance. Elle s'adapte au contenu, est facile à calculer et permet une réduction optimale de l'image à chaque itération. Cependant, l'encodage des seams peut rapidement devenir très lourd en raison de leurs formes complexes.

Suite à cette description des différentes méthodes de redimensionnement, nous avons fait le choix d'utiliser le « seam carving », en cherchant à réduire le coût d'encodage des seams.

1.4.2.3. Un encodeur vidéo

La vidéo réduite doit être encodée afin de réduire au mieux la redondance spatio-temporelle et le débit. Pour cela, on peut utiliser les anciens standards de compression tels que MPEG-1 et MPEG-2, le standard actuel H.264/AVC ou encore le nouveau standard HEVC. Puisque nous ne modifions pas le codeur, notre approche peut fonctionner avec chacun d'entre eux. Nous faisons cependant le choix d'utiliser le codeur H.264/AVC pour des raisons de performance face aux anciens standards MPEG-1 et MPEG-2 et pour des raisons de connaissance et d'implémentation par rapport au nouveau standard HEVC.

1.4.2.4. Une représentation peu coûteuse de l'information supprimée pour revenir aux dimensions d'origine

Comme il est expliqué précédemment, le seam carving a l'avantage d'être une méthode de redimensionnement qui s'adapte au contenu et qui donne un résultat de redimensionnement optimal pour chaque itération. Cependant, les seams ont des formes pouvant être complexes et donc difficiles à encoder. Afin de réduire le coût d'encodage de ces seams, seule une partie des informations sur les seams doit être transmise. On cherchera avant tout à définir ce qui est important de transmettre pour ne pas déformer les objets et bien les repositionner. En fonction de cela, il faudra mettre en place une modélisation des seams et une méthode d'encodage.

1.4.2.5. Les contributions de la thèse

Durant cette thèse, plusieurs problématiques ont dû être traitées. Au vu de l'importance de bien détecter les zones saillantes, des travaux ont été faits sur les cartes de saillance.

(1) Saliency : ST-RARE [Décombas 2013a]

(1) Une méthode de carte de saillance ST-RARE (Spatio-Temporal saliency based on RARE model) basée sur des informations spatio-temporelles a été proposée dans [Décombas 2013a]. (1a) Des informations de mouvement (direction, vitesse) ont été ajoutées au modèle de rareté spatiale (couleur, texture, luminance), (1b) le mouvement global a été supprimé afin d'améliorer la précision et la robustesse des résultats de saillance, (1c) un module de tracking permettant de combiner l'image courante avec l'image précédente a été rajouté, ceci afin d'augmenter la robustesse temporelle. Cependant, les résultats sont flous et pas directement liés aux objets.

(2) Saliency : STRAP [Riche 2014]

Il s'ensuit (2) [Riche 2014] une méthode de saillance spatio-temporelle basée sur la rareté dénommée STRAP (Spatio-Temporal Rarity-based algorithm with Priors) et incluant des à priori, a été développé pour de la modélisation du regard humain et de la détection d'objets. Dans ce papier, nous proposons (2a) une compensation temporelle du mouvement sur une fenêtre glissante

permettant de gérer à la fois les caméras fixes et les mobiles et fournissant (2b) des caractéristiques spatiales et temporelles sur la vidéo. (2c) Ces caractéristiques sont combinées avec un nouveau modèle basé sur la rareté et des à priori bas niveaux. (2d) Des à priori hauts niveaux sont combinés aux modèles de saillance afin d'améliorer les performances et (2e) un modèle de segmentation est utilisé afin d'avoir une approche orientée objet. Afin de valider les résultats, la base de données non compressée incluant des résultats de suivi de regards [Hadizadeh 2012] est complétée avec (2f) des masques binaires manuels afin d'évaluer la détection d'objets. Cette base de données a été choisie du fait qu'elle est sans artefacts de compression. Comme de nombreux modèles de saillance sont proposés et validés sur différentes bases de données (2g) nous avons comparé nos résultats en utilisant, en fonction des autres modèles, différentes informations hauts-niveaux. 7 modèles ont été choisis pour les évaluations avec 3 références (suivi du premier regard, suivi du second regard, masque binaire) et 4 différentes métriques.

(3) Compression vidéo par « seam carving » [Décombas 2011]

Afin de réduire le coût d'encodage des seams, différentes méthodes de modélisation ont été proposées. Dans [Décombas 2011], une modélisation par lignes clés a été proposée (3). Ceci a permis de (3a) présenter un premier schéma d'encodage par seam carving, (3b) de définir et identifier les zones importantes à l'aide de carte de saillance, (3c) d'encoder ces lignes clés où les seams se concentrent, (3d) de modifier la fonction d'énergie cumulative afin de contrôler la réinsertion des seams du côté du décodeur.

(4) (5) Compression vidéo par « seam carving » [Décombas 2012a] [Décombas 2012b]

Une autre modélisation a été proposée dans [Décombas 2012a] (4) et améliorée dans [Décombas 2012b] (5). Cette modélisation est basée sur un regroupement des seams, nous avons proposé (4a) une nouvelle fonction d'énergie prenant mieux en compte l'aspect temporel, (4b) une meilleure combinaison de la carte de saillance et du gradient, (4c) une identification des groupes de seams par un k-médian, (4d) une synthèse du fond par « shift-map », (5a) un rebouclage qui permet d'utiliser à l'encodeur les seams approximées du côté du décodeur, ceci afin de réduire les déformations géométriques.

(6) Compression vidéo par « seam carving » [Décombas 2014]

(6) La dernière modélisation proposée [Décombas 2014] va permettre (6a) de découper automatique et en fonction du contenu la séquence en GOP. Un meilleur taux de réduction spatiale pourra être ainsi obtenu. (6b) Un regroupement spatio-temporel basé sur une distance spatiale et une distance temporelle permettra de (6c) supprimer les seams isolés, améliorant ainsi les performances d'encodage. Les groupes les plus importants seront (6d) modélisés afin de limiter les distorsions géométriques et de diminuer la complexité du décodeur. (6e) Finalement, un nouvel encodage des seams est proposé et permettra de limiter le surcoût.

Les modélisations ont fait l'objet d'un brevet international [Décombas 2012d]

(7) Métrique de qualité orienté objet [Décombas 2012c]

Afin d'évaluer les performances, à la fois en terme de réduction de débit, mais aussi en terme de qualité sur les objets d'intérêt, une métrique [Décombas 2012c], SSIM-SIFT (7) a été proposé. Cette

métrique pleine référence permet d'évaluer les objets d'intérêt ayant subi des artefacts de compression et des déformations géométriques. Elle est basée sur (7a) une combinaison du SSIM et du SIFT. Cette métrique est (7b) insensible au background et à la synthèse de celui-ci, permet de (7c) mesurer les artefacts de compression de type H.264/AVC (SSIM_SIFT) et les déformations géométriques (Geometric_SIFT). Une validation subjective a été faite.

(8)(9) Résumé vidéo par seam carving [Décombas 2013b] et [Décombas 2013c]

Finalement, dans [Décombas 2013b] et [Décombas 2013c], nous avons aussi proposé (8)(9) une méthode de résumé vidéo par regroupement spatio-temporel de seams. Cette nouvelle méthode va permettre (8,9a) de déterminer le taux de réduction temporelle en fonction du contenu, (8,9b) supprimer les groupes de seams isolés créant des déformations géométriques, (8,9c) d'identifier des groupes spatio-temporels de seams suffisamment larges et (8,9d) d'approximer par des segments constants le nombre de seams dans chaque groupe tout en gardant le nombre total de seams constant. Ceci permet d'éviter les artefacts géométriques sur les objets d'intérêt.

Table of contents

2. Introduction

2.1.Context

2.1.1. Introduction to defense and security applications

In the past few years, countries and companies have decided to improve their ways of protecting citizens. Solutions in video surveillance have emerged all over the world. Developed countries like the United States or United Kingdom do have a very thorough policy on that particular matter. Developing countries also consider video protection as a mean of lowering criminality. For example, the capital of Mexico was the most dangerous city of the country and the government decided to put in place a 3-year project of video surveillance of approximately 460 million dollars. That project, a partnership between Thales, Mexico and Telmex, was set up to allow the city to manage a greater number of risks, like delinquency, terrorism, attacks against strategic sites or natural risks, and so 8080 cameras and sensors were deployed.

Big companies see this solution of video protection as a good way of protecting their assets and employees. It allows them to respond faster to thievery, attacks or arson. In the last 15 years, groups like Thales have developed solutions in civil and urban safety for countries in order to improve efficiency of first response agents and lower costs of surveillance. It includes the monitoring of public transportation. For instance, the city of Paris has a subway system comprised of 14 lines and 300 stations on a total of 213 km of train sets [RATP]. So protecting 1.4 billion users against aggressions, pickpockets and fraud is not humanely possible. In order to attempt to resolve these specific problems, the RATP, agency responsible of public transit, has hence developed solutions in video surveillance by deploying 8200 cameras in its subway system and 18,000 on its buses [Parisien 2012].

Video protection solutions also allow to manage areas too dangerous for human beings. They can be applied to military camps possibly protected by deployed sensors, data retrieval done by flying or unmanned ground drones, etc. All these solutions are designed to improve responsiveness and coordination of actions in case of crises like a terrorist attack or a natural disaster. Growing demand of these solutions has allowed to lower the cost of sensors and to better integrate cameras in our environment.

United Kingdom has now more than 1.85 million cameras [Reeve 2011], [Dailymail 2011], and the United States have 30 million [Popularmechanics 2009], [Vlahos 2008]. The French government has launched a large strategy on video protection in 2010 [Figaro 2010] and authorized in 2013 video reporting for petty offenses [Lepoint 2013].

More customized solutions are being developed by iwatchlife [Iwatchlife 2013] which offers a solution in video protection against break-ins and other abnormal events, and all the data is sent to your cell phone. Other companies like SecuriteEnEntreprise offer temporary protection of sites, as festivals, deliveries or construction sites, for which video protection solutions can be very useful [Securiteentreprise].

However, almost all the proposed solutions included the use of traditional video encoders that are mass-produced. Mostly used for video transmission or movie recording, these tools are designed to maintain the overall quality of image and obtain a pleasant rendering. Image quality is not essential in professional applications, where detection, understanding of events and reliability in all conditions are main priorities.

2.1.2. Requirements

2.1.2.1. Sensors in large quantity and/or of good quality

Prices of IP cameras have dropped considerably mainly due to their growing demand [Mantratec]. Densification and increase of network capacities have allowed to connect all cameras to networks. In wire systems, transmission capacities were of 10 Gbps [Landa 2006] [Carte_Reseau] and concurrently different types of wireless systems were being designed. For instance, Bluetooth makes it possible for two objects to communicate on short distance, WIFI allows a computer or a smartphone to connect to public or private networks, and other mobiles solutions also exist (edge, 3G, 4G). All those technologies permit interconnection of devices with always increasing transmission capacities.

Network capacity, definition of screens and sensors and the willingness to obtain images of a constantly improved quality paved the way for the transition of image format 640x480 Video Graphic Array (VGA) in 1987 [VGA] to other ones like full HD (1920 x 1080) or the newest 4K (3840 × 2160).

Nowadays videos require higher bitrate and can be transmitted in locations with a well-developed network. However, the available bitrate can vary widely depending on the location. Robustification of sensors and increase of battery capacities have allowed to install high definition cameras on vehicles, people, drones or directly on the ground.

2.1.2.2. Very low or too costly available bitrate

As seen previously, network capacities are becoming more important, but not necessarily well distributed. Solutions in telephone network of fourth generation (4G), individual network (WIFI) and optical fiber are available in big cities, but in the country side of developed countries, we come across very soon with low capacity bitrates. As for developing countries, since wire systems are too costly to put in place, wireless systems are preferred, which means GSM-based solutions are present almost everywhere in the world provided that there is a certain human density. However, in areas of very low human density, like deserts, large forests, big lakes and seas, and even oceans, there is no mean of transmitting data.

The presented solutions are proposed by public suppliers and are available to everyone. But, for defense and security applications, one must have its own network in order to control data transmission from one end of the chain to the other and hence ensure data safety. Also, defense and security solutions have to be reliable and working in any location. In this case, public network are an option only if they can be used in any given situation and if transmitted data does not need to be highly secured. Defense and security applications can also use two other types of transmission: by radio which has a low bitrate or by satellite which is very costly.

2.1.2.3. Hard to access or dangerous locations

Professionals have to be able to work in all given situation and location.

Certain locations are hard to access and have low available bitrates, as high mountain areas. For instance, the Institut de Formation et de Recherche en Médecine de Montagne (IFREMMONT) [Iffremont] seeks to develop e-medecine solutions for purposes of remote medical assistance. Hence video transmission is likely to be used, allowing a doctor to analyze the situation and give the necessary medical assistance while waiting for first aid to come.

Sea, air and space are also hard to access and very vast locations, thus difficult to link to networks. Deploying human protection to monitor deserts or frontiers of a country is nearly impossible to achieve.

Last but not least, certain locations can also be dangerous: war zones or areas affected by a natural or human disaster. In those particular cases, it is possible to deploy one or more intervention groups but they always will be in danger. They could be under attack in war zones and subject to hard climate conditions in areas affected by a natural disaster. However, in all these cases, solutions in video protection can lower risks but only if the data necessary for decision-making is transmitted.

2.1.2.4. Need to retrieve data necessary for decision-making

In presence of more sensors, the need to be reliable and working in all given situation and location, and the difficulty to transmit data, we have to define what is important to transmit and how the data will be used.

At first, data is transmitted to a decision-making center. It has to help analyze the problem in order to understand the situation and issues. After which, the control center has to be able to make decisions and give orders and guidance in order to change or improve the situation in the field. Decision-making at the control center allows to reduce costs of an operation, for instance, by avoiding using of a helicopter for a rescue in high mountain area or of a drone for a second time for it to retrieve data over enemy lines in order to accomplish its first mission. It also allows to reduce risks for humans by avoiding sending an intervention group in a dangerous location or by better coordination of a group in the field. With those considerations in mind, a few scenarios of applications will be presented with further details.

2.1.3. Scenarios

2.1.3.1. Deployed sensors

The first scenario to be presented includes deployed sensors.

This type of sensor allows to send a warning in an area not monitored by a person. These areas can be dangerous, suspicious or very vast.

In this scenario, video sensors are usually combined with others, like seismic, laser and audio sensors.

There are two reasons for this. Firstly, sensors have to be able to send a warning and will be more robust if they use different stimuli. Secondly, video sensors need large amounts of energy, and if they are deployed for many months, they need a very efficient energy management system. The simplest solution is to leave them inactive until other sensors receive a stimulus. In addition to the energy constraint, there is the human constraint. Indeed, the time needed for implementation and deployment has to be short, because a sensor is by definition deployed in an area where a human being cannot stay long. Also, sensors have to be robust under all scenarios and conditions and quick to set up.

Here is an example of scenario: the police suspects that illegal activities are taken place in a particular location, which is in an area with poor network. Sending an intervention group is not an option for reasons of economy and discretion. Installing a camera with a battery is feasible, but retrieving all the data is not. Hence it is hard to know when to pick up the sensor without being seen.

On the other hand, permanent transmission would make the sensor easy to detect and take large amounts of energy. A solution based on low bitrate content-aware video transmission would solve the problem.

2.1.3.2. Drones

The second scenario includes drones.

Drones are unmanned vehicles, aerial or ground, that are used for surveillance and mapping of dangerous areas. Video sensors used are often of high definition in order to retrieve as much data needed for analysis as possible. Drones go on reconnaissance in suspicious areas with a possible low bitrate network. However, a drone could reenter a better covered area to send out a part of the data retrieved and fly back.

Here is an example: in a normal case, a drone follows its flight plan and, once it returns to base, the data retrieved is analyzed. If after analysis, it is recognized that some data is missing, a new flight plan has to be drawn. This can be costly and risky because the drone could have been detected by the enemy or it could fly over a turbulent area where it could be destroyed. With a solution in transmission that is adapted to content and at low bitrate, the drone could send out a sequence on the area of interest, which would allow the command center to analyze that data sample and correct the flight plan if needed.

2.1.3.3. Onboard cameras

The last scenario to be presented includes onboard cameras, which means they are placed on a vehicle or a human being.

In concept, it is quite similar to the drone, but it much differs in its applications. Onboard cameras can be used so their bearer is able to better detect activities in his surroundings, even to transmit data to a command center that will view it and contribute to decision-making. Onboard cameras also allow ex-post analysis of interventions, which promotes better understanding of events and correction of certain actions in order to improve the next missions.

Real-time or slightly delayed transmission can improve actions during an intervention. Professionals (policemen, firefighters or infantrymen) are being equipped with cameras while on intervention. The equipment used can be numerous high resolution sensors (visual, infrared or thermal imaging camera), which can provide large amounts of data. The network might be incapable to manage it, and this incapacity might become a constraint when transmission to command center will be needed. Another constraint often occurs when several onboard cameras bearers are too close to each other, use the same transmission antennas and saturate the network. As mentioned before, for defense and security applications it is necessary to avoid depending on a public supplier and to ensure efficiency in any given situation. In that case, transmission tools are necessarily onboard and do not have a large capacity of transmission. For example, a team of firefighters are entering in a building on fire. Each firefighter has a thermal camera and a visual camera on him. The command center will send its team while listening and analyzing all radio conversations. Content-aware video transmission would allow the command center to visualize automatically all actions and observe problems encountered on the field in sensitive areas. The command center will then ensure better leadership, organization and safety of its team during the intervention.

In the three scenarios, it is not necessary to access all the data, since only a part of it is useful and used in decision-making.

Currently, the usual solution for defense and security applications is to use H.264/AVC which is a good encoder for general public applications but is not necessary designed for defense and security ones.

2.2.Encoding

2.2.1. Advantages of encoding for general public applications

For the purposes of evaluating the advantages of encoding, we will take the example of a 90 minutes full HD movie that has 25 images per second. So the size of the file will be calculated by the following formula:

Nb_Images_per_Sec x Nb_of_Sec x Nb_of_Channels x Spatial_ Dim x Dynamic = 25 x 90 x 60 x 3 x 1920 x 1080 x 1 = 838,808 GB.

Without encoding, a simple movie is difficult to save and to transmit. For commercial and technical reasons, movies have to be in a size inferior to the physical support used. A CD has a 700 MB capacity, a DVD has capacity of 4,37 GB to 8,54 GB and a Blu-ray, from 7 to 100 GB. Even with the best methods of physical storage, we would need 8 Blu-ray for a 90 minute movie, which would imply that the user would have to change disk every 12 minutes.

During encoding, the objective is to maintain the overall quality of the movie and the best possible quality depending on the support or transmission capacity. To measure the overall quality of the image, objective metrics as the Peak Signal to Noise Ratio (PSNR) or the Structural SIMilarity (SSIM) [Wang.Z 2004] were established. Encoding methods aim at maximizing quality according to available bitrate.

2.2.2. Data reducing methods

Video sequences contain a great quantity of redundant data, and compression technologies are based on a strong correlation between pixels, as much spatial wise where adjacent pixels are similar, as time wise where pixels of past and future images are also very close.
Encoding methods are looking to use a combination of Intra-Frame and Inter-Frame coding. Intra-Frame coding would use spatial correlation between pixels while Inter-Frame coding would use temporal correlation.

2.2.2.1. Subsampling and interpolation

A simple method to reduce the quantity of data consists in taking a subsample of the source, which would reduce spatial dimensions contained in the video sequence and thus the number of pixels to code. It is also possible to subsample the temporal axis by reducing the number of images per second. To obtain a final result with spatial dimensions and frequency identical to the source, it is necessary to execute an interpolation at the decoder which permits to synthesize deleted parts.

2.2.2.2. Intra Coding (I-frame)

I-frames take advantage of spatial redundancy and do not depend on data from other frames. I-frames are divided into pixels blocks. The data in each block is transformed from the spatial domain to the frequency domain by a DCT. The image can be simplified by quantizing the coefficients in the

frequency domain. High frequency components are usually put to zero which causes a lost in the nuance of the image. The quantized coefficients are then compressed by using run length codes and Huffman coding.

2.2.2.3. Inter encoding

The P-frame provides more compression than I-frame by taking advantage of the previous I- or P-frames, defined as reference frame. The frame is divided into 16x16 macroblocks and for each macroblock, the best matching in the reference frame is searched. The offset is encoded as a motion vector. The matching is in general not perfect, so the differences of all corresponding pixels of the two macroblocks are encoded and represent the residual. When no matching is found, the macroblock is treated like an I-frame macroblock. The processing of B-frames is similar to the P-frames but in the B-frames, the next and the previous frame are used as reference frame. The B-frames are never used as reference frames.

2.2.2.4. Variable length coding

Once predictions are done, symbols have to be used to transmit a certain amount of bits. The simplest method is the fixed length coding, where the number of bits stays unchanged, regardless of the symbol. That number of bits is defined by the dynamics of symbols. For example, a pixel can have a value between 0 and 255 for each of its RGB components. We thus need 2^8 = 256 values to represent symbols, which are 8 bits by component.

Variable length coding is used to associate a symbol of the source to a variable number of bits and allows the source to be compressed and decompressed without any error. This makes the coding lossless. In fact, a low number of bits are associated to the most frequent symbols and, conversely, a high number of bits to the less used symbols. Huffman [Huffman 1952], Lempel-Ziv [Welch 1984] and arithmetic coding [Witten 1987] follow this principle.

2.2.3. Advantages of encoding for defense and security applications and conclusion

For defense and security applications, the purpose of encoding is ensuring a link between the source and the recipient under any condition. Loss of that link may cause severe consequences, both human (incapacity to manage a field team in a dangerous area) and economic (incapacity to modify a drone's flight plan). This is where lies the main difference with entertainment applications for which data loss is annoying but without vital consequences. In defense and security applications, data to be transmitted does not have to be of great quality, unlike entertainment applications, but is destined to maintain the semantic for purposes of analysis and decision-making.

Traditional encoders have been designed to meet public needs, meaning to obtain an image of better quality on supports with ever growing capacities.

Traditional encoders are not directly adapted to defense and security applications and may work incorrectly when the bitrate is too low. In fact, if we wish to maintain a good overall quality at low bitrate, the whole image will be deteriorated and the scene's semantic lost. There is no willingness to adapt to content and to do semantic encoding.

2.3. Objectives

In presence of low bitrate, it is hard to keep all the data. We have to decide what to keep and how to reduce the quantity of data in a smart way. We must establish a certain number of essentials goals depending of constraints and defined scenarios.

2.3.1. Staying compatible with existing standards

Since the solution we wish to propose should be designed to be integrated to numerous existent products, we have to modify the less possible the existing solutions for economic and compatibility reasons. By this way, the traditional coder can be changed without any difficulties. We decided on a process that can be applied in upstream and downstream of traditional encoding.

2.3.2. Defining what is necessary to interpret video footage

As the desired encoder has a main objective of maintaining both semantic and video interpretability, it has to be able to detect salient areas. Once these areas detected, we have to allocate the bitrate and obtain a good-quality rendering of salient objects in order to identify and to define them. It is also important, for understanding and interpretation purposes, to clearly identify the position of objects and their relative motion while keeping enough data on the background to define context and environment. The user must always be critical about the applied processes by leaving them easily identifiable. In the case of bad detection of a salient object, this precaution will allow the user to take into account any errors made by the encoder and take the necessary steps to correct them by asking the source to resend the whole frame without any treatment for example.

2.3.3. Having flexible transmitted data and working at low bitrate

Once saliency areas are detected, we have to define a way to automatically separate them from the background.

When this is accomplished, we have to find a process to best concentrate the pertinent data and to delete less pertinent data, which is usually concerning the background. We also have to remember that some data has to be kept in order to know the context. Reduction of non-pertinent data gives us the necessary flexibility in regards to what we wish to transmit, for example a better quality for saliency objects, another sequence, or more data on the context. The data to be transmitted has to stay easily scalable so the user can readily distribute the available bitrate according to his needs. If he ever needs to transmit more data, it is necessary to compact the data as much as possible to best limit cost overrun.

2.3.4. Being measurable

We have to be able to measure to the extent possible the interpretability of video sequences. We can get around the background and measure separately geometrical distortions that objects and their position are subject to. Measuring the quality of salient objects remains essential in order to evaluate if they contain enough interpretable details and if they can still be defined.

2.3.5. Working in different environments

Knowing all application scenarios, we have to keep in mind that the encoder we propose has to be easy and reliable to implement. Settings have to be at a minimum, if not adaptive to the content. The encoder has to be able to work in any context, because of the lack of knowledge on both scenarios and environment. It could reveal itself dangerous to enter a definition of scenario, context or

environment that has been previously predicted or established, because there is always the risk of being confronted to a new or unpredictable situation.

2.3.6. Low bitrate compression by saliency area

Without any predefined scenario or particular equipment in mind, we set a theoretical objective to transmit video between 300 to 600 Kbit/sec. The idea is to identify the feasibility of this approach for deployed sensors, drones or onboard cameras.

No constraints were done on complexity, energy consumption, camera definition or sensor type (infrared, visible, etc.). In this case, we propose a low bitrate compression method by saliency area that responds to numerous problems and conditions associated with each scenario. This will help us achieve a feasibility study.

2.4. Proposed solution and contributions

2.4.1. The principle

We are proposing here to compress video by resizing it.

The idea is to delete insignificant part of data contained in the video to reduce the bitrate.

At first, we apply spatial resizing on the original video. Then the reduced video is encoded with a traditional encoder and data allowing to return to original spatial dimensions is also encoded. In regards to the encoder, the reduced video and all data concerning spatial resizing are decoded and combined in order to return to a video with its original dimensions. Deleted areas can be synthesized if it contributes to understand the scene. Figure 3 summarizes the principle of this encoding process.

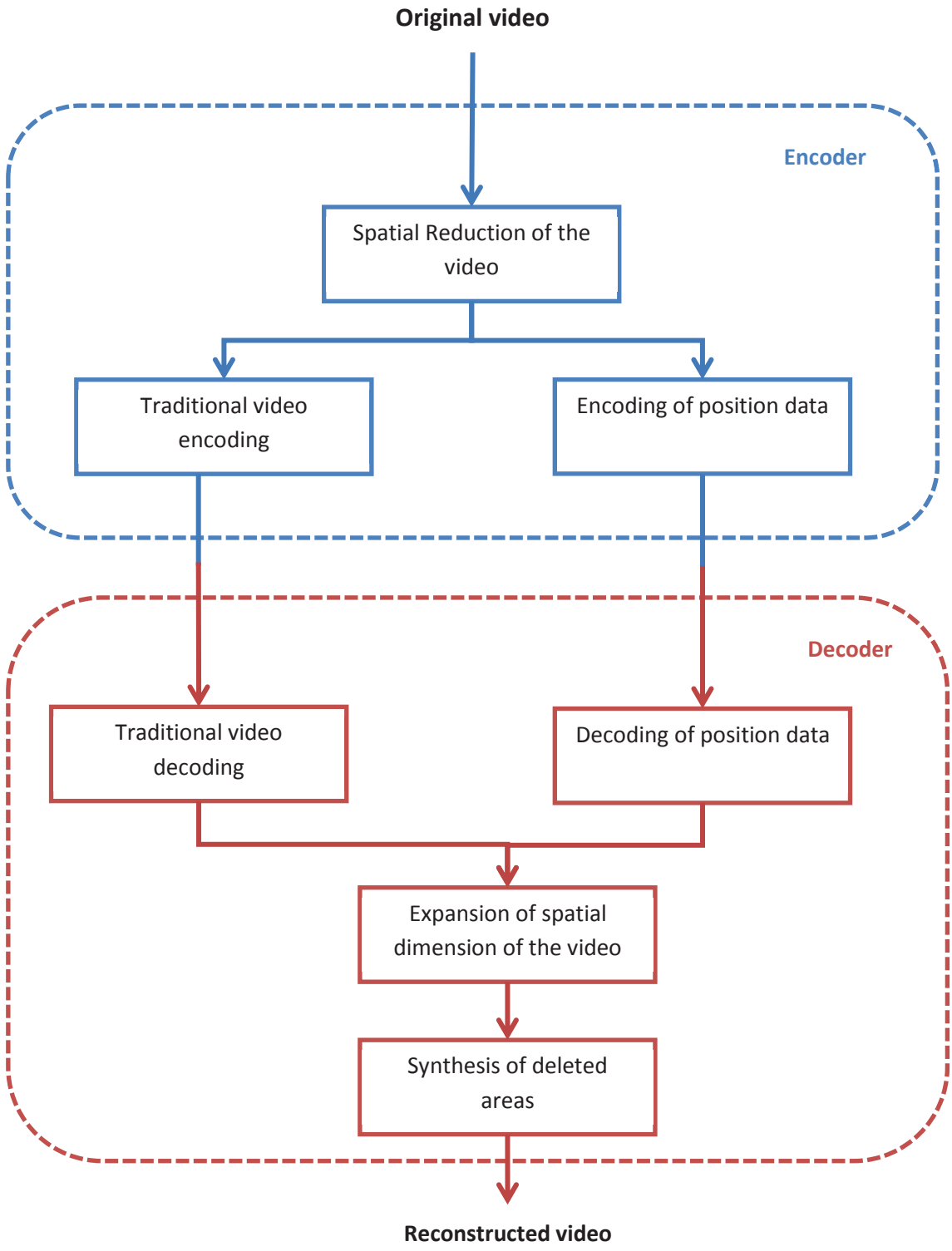


Figure 3 Diagram on video compression through resizing

2.4.2. Useful tools

In order to realize this encoder, several useful tools are necessary

2.4.2.1. A tool to detect saliency areas

A tool to detect saliency areas is needed. There are two main approaches: background extraction and saliency maps.

Background extraction aims at modeling the background and at defining salient objects that are not the background. This approach works well when the camera is fixed and permits a precise clipping of salient objects most of the times. However, the result is binary: is the background or not. To know what constitutes the background, it is often necessary to have a learning phase. These approaches can be sensitive to brightness changes and do not work if the background changes too fast.

Saliency maps [Koch 1985] seek to model the human eye to predict where a user would look. They work in most cases and need little initial conditions. In conclusion, with these approaches, each pixel has the probability of being looked at. The disadvantage of these methods is that results are often blurred and the borders of saliency area are not linked to the salient object's one.

In conclusion, as the proposed solution should work for drones, onboard cameras or deployed sensors and as there is no assumption about the environment, the tool should automatically define what is salient without needing a learning phase. Based on these considerations, saliency maps are chosen and the solution that will be design should give a sufficiently accurate result to preserve the objects.

2.4.2.2. Method of spatial resizing

As soon as salient areas are known, less pertinent data is deleted with a spatial resizing method. There are at least five of them: resizing, cropping, warping, data pruning and seam carving.

Resizing is a method that evenly reduces a video with a pyramidal approach. Each reduced pixel on the *n* scale is associated with the average value of a set of pixels on *n-1* scale. The interest of this method relies in its simplicity of application. However it is not possible to make it content-aware and it reduces also salient objects. Hence details on them are lost. Even though additional data on salient object is sent, this approach is difficult to reverse.

Cropping is a method that deletes data on the borders of the frame. It is also easily applied and there is little data to transmit in order to return to original dimensions. However, this approach works well only if salient objects are in the frame edges and it is not possible to make it content-aware.

Warping is based on the gridding of the frame, and each square is resized in function of its saliency. This resizing approach has the advantage of being content-aware and of less reducing salient areas. However, the approach is computationally intensive and the resizing information can be difficult to model.

Data pruning consists in removing straight lines inside the frame according to their saliency. This content-aware approach is easy to compute and facilitates encoding of resizing data. However, it quickly reveals its limits: salient areas can have complex shapes and straight lines will touch salient objects.

Finally, the last method of resizing is called seam carving. This approach deletes seams inside the frame in function of their saliency. It is content-aware, easily computable and allows an optimal reduction of the image at each repetition. However, the coding of seams can be very heavy due to the complexity of their shape.

Following the description of different resizing methods, we chose to use seam carving while seeking to reduce the cost of encoding seams. The seam carving will be applied and adapted for video and

works on the shape complexity of the seams will be done. The temporal aspect will be an important point to take into account.

2.4.2.3. A video encoder

A resized video has to be encoded in order to reduce spatio-temporal redundancy and bitrate. For this purpose, legacy standards, as MPEG-1 and MPEG-2, the actual state of the art standard H.264/AVC or even the upcoming HEVC can be used. Since we do not to modify the encoder, our approach can work with all these standards. We chose to use the H.264/AVC encoder for reasons of performance, compared to MPEG-1 and MPEG-2, and for reasons of knowledge and implementation, compared to HEVC.

2.4.2.4. A low cost representation of deleted data to return to original dimensions

As previously explained, seam carving has the advantage to be a content-aware resizing method that gives an optimal resizing result for each repetition. However, seams can have complex shapes difficult to encode. In order to reduce the cost of encoding these seams, only a part of the data on seams has to be transmitted. We primarily seek to define what is important to transmit to prevent any distortion of objects and correctly repositioning them. According to that, we will have to compute seams and create an encoding method.

2.4.2.5. Contributions of this thesis

During this thesis, different problematic have been seen. As it is indispensable to well detect salient objects, works on saliency maps has been done. The two contributions on saliency maps are detailed in chapter 4.

(1) Saliency: ST-RARE [Décombas 2013a]

A method of saliency maps, called ST-RARE (Spatio-Temporal saliency based on RARE model) (1), based on spatio-temporal information has been proposed in [Décombas 2013a]. (1a) Temporal information (direction and speed) has been added to spatial ones (color, texture, luminance) in input, (1b) the global movement of the camera has been suppressed in order to improve the precision and the robustness of the saliency results, (1c) a tracking module allowing to combine the current saliency map with the previous one has been used in order to improve the temporal robustness. Nevertheless, results are blurred and not directly linked with the objects.

(2) Saliency: STRAP [Riche 2014]

Due to that, in [Riche 2014] (2) a Spatio-Temporal Rarity-based algorithm with Priors (STRAP) for human fixations prediction and objects detection in videos has been developed. In this paper we propose, (2a) a temporal compensation of the movement on a sliding windows allowing to manage both static and moving cameras and (2b) giving more robust spatial and temporal features. (2c) These features are combined together with a new model based on rarity and low priors. (2d) High priors information are combined to the salient models to increase the performance and (2e) a segmentation model is used to have a more object based approach. To validate the results, the raw format database with eyes tracking results from [Hadizadeh 2012] is completed with (2f) manual binary masks to evaluate the detection of the salient objects. This database has been chosen due to the fact that it is free of artifacts. As lots of saliency models are proposed and are validated on different database, (2g) we compare our results with different high priors information to be

consistent with the other models. 7 models are chosen for the evaluation on the 3 references (eyes tracking first view, eyes tracking second view, binary mask) with 4 different metrics.

After having defined what is important with the saliency models, seam carving will be used to suppress the less important regions and by this way, the quantity of information to encode will be reduced. 4 models have been realized during this thesis and are presented in chapter 5.

(3) Video compression by seam carving [Décombas 2011]

In order to reduce the encoding cost of the seams, different modeling methods and seams encoder has been tested. In [Décombas 2011], a seam modeling based on key lines has been proposed (3). This approach allows to (3a) present a video encoder based on seam carving, (3b) to identify important lines defined as the areas where the seams are concentrated. These areas are between salient objects and their encodings allow a good repositioning of the objects. (3c) An encoder of these lines is presented and (3d) a modification of cumulative energy maps in order to control the seams reinsertion at decoder side is proposed.

(4), (5) Video compression by seam carving [Décombas 2012a], [Décombas 2012b].

Another model is presented in [Décombas 2012a] (4) and improved in [Décombas 2012b] (5). This model is based on seams clustering. A new energy function is presented (4a) to better take into account the temporal aspect, (4b) a better combining of the saliency maps and the gradient, (4c) an identification of groups of seams by k-median, (4d) a seam texture synthesis based on shift-map are described. In [Décombas 2012b] (5a), the seams are re-synthesized similarly to the process carried out at the decoder side. The resulting information is used to define a new reduced sequence. This closed loop process leads to less geometric deformation when the initial seams are spatially scattered in the sequence.

(6) Video compression by seam carving [Décombas 2014]

The contributions (6) proposed in [Décombas 2014] is (6a) an algorithm that automatically cuts the sequence into GOPs, depending on the content, (6b) a spatio-temporal seam clustering method, based on spatial and temporal distances, (6c) an isolated seam discarding technique, improving the seam encoding, (6d) a new seam modeling, avoiding geometric distortion and resulting in a better control of the seam shapes at the decoder without saliency map and (6e) a new encoder that reduces the number of bits to transmit.

The different models have been also described in an international patent [Décombas 2012d].

After having defined the important parts with the saliency models and suppressed the less important regions with the seam carving, it was necessary to evaluate the quality of the salient objects. This contribution is presented in section 6.4.

(7) Object based quality metric [Décombas 2012c]

To evaluate the performance of the encoders in function of bitrate saved and the quality of the objects of interest, rate distortion assessment are done with a proposed metric in [Décombas 2012c] called SSIM_SIFT (7). This full reference metric allows to evaluate the quality of the object of interest having encoding artifacts and geometric distortions. (7a) It is based on a combining of SSIM and SIFT.

This metric has the advantage (7b) to be insensible at the background suppression and synthesis and allows to (7c) measure compression artifacts due to traditional encoders like H.264/AVC (SSIM_SIFT) and the geometric deformation of the objects (Geometric_SIFT). Subjective validation has been done showing its efficiency.

We noticed during this thesis that the approach to do video compression by resizing can be transposed to do video summary. It has been tested and 2 contributions have been done. This is detailed in chapter 8.

(8) (9) Video summary by seam carving [Décombas 2013b] and [Décombas 2013c].

Finally, in [Décombas 2013b] and [Décombas 2013c] a method of Spatio-temporal grouping with constraint for seam carving in video summary application is presented. In these papers, We propose (8,9 a) a way to do an efficient spatio-temporal grouping that allow us (8,9 b) to determine a temporal rate of reduction in function of the content, (8,9 c) to suppress the group of isolated seams, (8,9 d) to identify sufficiently large spatio-temporal groups of seams, and (8,9 e) to approximate by constant segments the number of seams for each group, while keeping the total sum of seams constant. The proposed method avoids geometric deformations of the salient objects and anachronisms. In addition, the summary has the same length on all the lines. Our constraint enables more flexibility in the seam carving process, with seams that can better adapt to the content. At the same time, a better rate of temporal reduction can be achieved while preserving salient objects.

3. State of the art

3.1.Saliency maps

3.1.1. Introduction

To understand saliency maps, it is important to define visual attention. It is the natural capacity to selectively focus on part of the incoming stimuli, discarding less important information. This ability was first develop by human being to survive because it is easy to imagine that if there is a predator in our environment, or foods are available in trees, we have to focus on that. Our capacity to identify and process the most important information as quickly as possible is essential. This capacity allowed us to better understand how the things work, to learn and evolve. Visual attention has been studied in different domains, like in psychology, neuroscience and engineering. The saliency is the way to model the attention in computer science. There are two approaches, the first one is a bottom up approach and the second one is the top down. These two approaches can be combined by including top down information in bottom up models.

3.1.2. Saliency model based on bottom up approach

The bottom up approach tries to extract the surprising, rare, novel information. The idea is to find the most pertinent information and to suppress the redundant one. This approach is based on features extracted from the signal such as luminance, color, texture, intensity of the movement, direction of the movement, symmetry, etc. On these features, the approach models our capacity to find in involuntary manner the most important stimuli. By this way, the salient parts are highlighted. This approach is based on a well understood phenomenon that is an advantage for the modeling. First, a study of proposed methods on still images is done, and then the case of video will be examined.

3.1.2.1. Case of the still images

This study has been, since a few years, very popular and lots of models are proposed each year. They are all based on the same idea, and differ in terms of implementations and technical approaches. In order to better see the main approaches, a classification is necessary. But it is not easy because models depend sometimes on the context, have different resolutions and can use idea from different categories. The three most known classifications will be present here. The one proposed by Judd in her thesis [Judd 2011] is “biologically-driven”, “mathematically based” and “top down information” and is linked with the research laboratories. Another one proposed by Borji sorts the models based on their mechanism to obtain saliency maps and are classified in a computational perspective [Borji 2013]. Categories like spectral analysis approach, “cognitive approach”, “Bayesian models”, “information theoretic”, can be cited. The classification of Mancas [Mancas 2012] is different and has the advantage to be more linked on the events and the way to process them. A classification in three categories: “pixel’s surrounding”, “the whole image”, “normality” is proposed.

In the “pixel’s surrounding” category, a pixel or a patch is compared with its surrounding to find some discontinuities. In 1985, Kick and Ullman, based on a biological motivation, introduce this field [Koch 1985]. The principle is to compute visual features at different scales in parallel, applies a center surround inhibition, combines the multiscale maps into features maps and then fused all the feature maps together to obtain a single saliency map. The first implementation has been done by Itti [Itti 1998] and it is decomposed in three main steps: The first step is the selection of features (color,

intensity, orientation) at different scales. Then at a second step, the center surround inhibition will give more importance for high contrast and lower importance for low contrast. Finally a fusion or combination step, where first each feature maps are combined together and then the feature maps are linearly combined together to obtain an inter-features map which is the saliency map. Itti tried different combination approaches and an efficient one was to give more importance for the maps having values much higher than the mean one. By this way, global information to the local one is added. Figure 4 illustrates the three main steps.

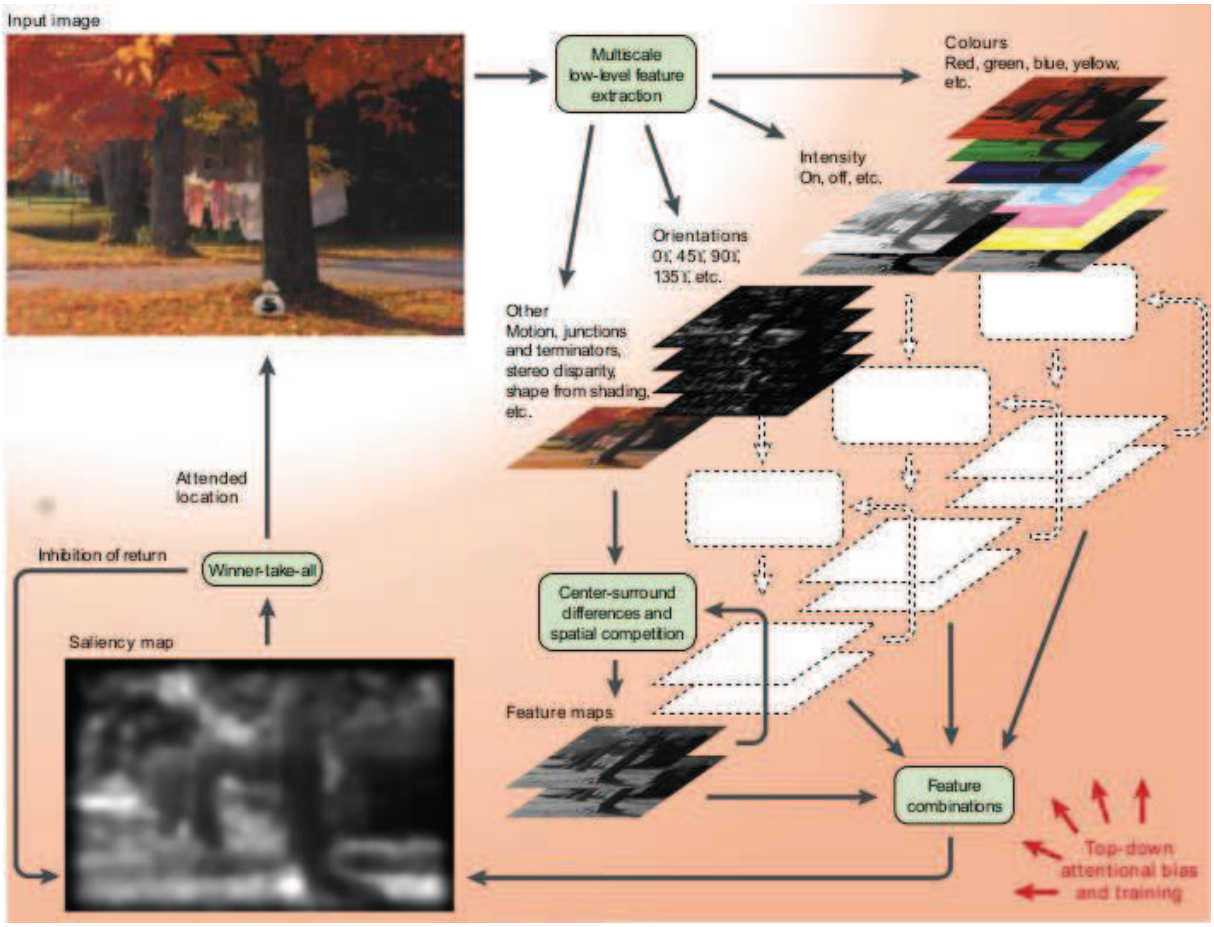


Figure 4 Model of Itti *et al.* [Itti 1998] in three steps: center surround difference, features maps and fusion into saliency map.

Privitera and Stark [Privitera 2000] improve the previous model by adding in input a symmetry feature. Valenti *et al.* [Valenti 2009] add curvedness information in input. Le Meur *et al*. [Le Meur 2006] improve the model with a more biological approach and add contrast sensitivity functions, visual masking, perceptual decomposition and center surround interactions. Another efficient and popular approach from Harel *et al.* [Harel 2006] is the Graph Based Visual Saliency (GBVS) model and differs from Itti during the fusion step. An activation map before the combining steps is computed that will give more or less importance to some areas. The algorithm builds a fully connected graph over all grid locations of each feature map and assigns weights between nodes that are inversely proportional to the similarity of feature values and their spatial distance. A center Gaussian is used to take advantage of the center bias and improve the results.

In “the whole image” category, the entire image is used. Pixels or patches of pixels are compared with pixels or patches of pixels from other locations in the image. The process can be divided into two steps: first the local features are computed in parallel in a given image, then the second step is to compute the likeliness of the pixels or patches of pixels. This kind of approach is called “self-resemblance”.

Seo and Milanfar [Seo 2009] use as feature local regression kernels and then kernel density estimation that estimates the distribution of the features in the patch is applied. We remind that in statistics, the kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. Their model has the advantage to be robust to noise and other systemic perturbation.

In [Culibrk 2010], Culibrk et al. proposed the use of a multi-scale background modelling and foreground segmentation approach, as an efficient saliency model driven by both motion and simple static cues. The model employs the principles of multi-scale processing, cross-scale motion consistency, outlier detection and temporal coherence.

Mancas proposes in [Mancas 2007a] and [Mancas 2009] a way to detect locally contrasted and globally rare areas that are considered as the salient parts.

Riche [Riche 2012b], [Riche 2013] continues on the work of Mancas to propose a bottom-up saliency model based on the idea that locally contrasted and globally rare features are salient. Low-level features like luminance and chrominance is combined with medium level features like orientations to obtain saliency maps. Their model has been compared with different models and different metrics on three databases of 120 images.

Stentiford [Stentiford 2001] starts from random patches and measures the quantity of similar patches in the entire image. Fewer patches mean the neighborhood is rare and salient. There is no need of feature extraction in this approach because it is directly included in the patches comparison.

Olivia *et al.* [Oliva 2003] define the saliency as the inverse likelihood of the features at each location. The features are extracted by using a steerable pyramid with 4 scales and 4 orientations and the likelihood is compute by using a Gaussian probability all over the image.

Boiman and Irani [Boiman 2007] compare not only the patches between them but also their relative positions.

A well-known model is the model proposed by Bruce and Tsotsos [Bruce 2005]. Their model is based on the principle of maximizing information sampled from a scene. This bottom-up model overt attention uses Shannon’s self-information measure and is achieved in a neural circuit. Random patches are projected on a news basis and used to weight the patches in the image. Figure 5 illustrates their approach.

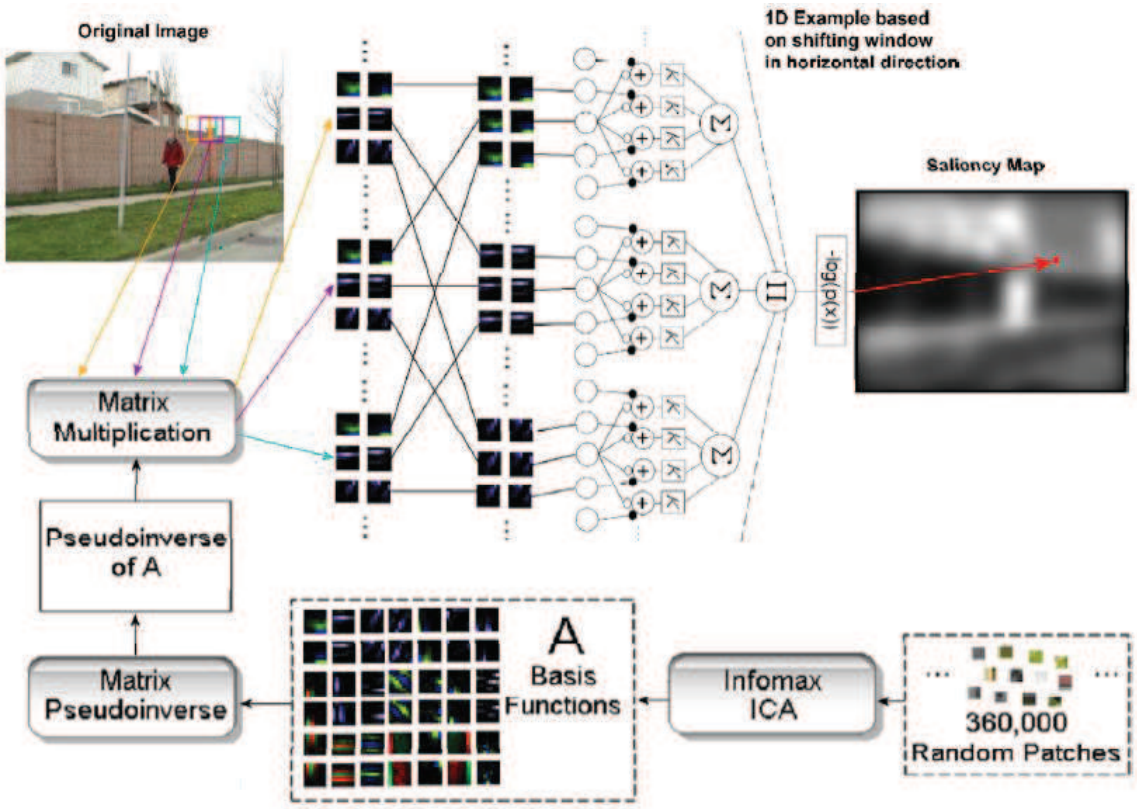


Figure 5 Framework of the approach from Bruce &Tsotsos [Bruce 2005].

Rathu and Heikkilä [Rahtu 2009] propose a method based on sliding windows where the distribution of an inside windows is compared with the distribution of the area surrounding the sliding windows. If the distribution are different, the inside windows have a high value, and if not, the value will be low. This process is applied at different scale to manage textures and salient object size. This method requires no training and no additional segmentation algorithms. Their works are illustrated for background subtraction.

Goferman *et al.* [Goferman 2012] build a content-aware saliency detection based on four principles of human visual attention. They use local low levels information like contrast and colors, global information that suppress frequently occurring features and maintain the rare ones, visual organization rules which define that visual forms may possess one or several centers of gravity about which the form is organize and finally high level information like human face detection.

The last category is the "normality" and the idea is to model what the things should be and by this way, the things that are not like they should be are surprising and salient. Achanta *et al.* [Achanta 2009a] develop a simple model where, in the CIE Lab color space, the Euclidean distance between a Gaussian filtered image and the mean value of the image is computed. Itti and Baldi [Itti 2006] define the concept of surprise by a formal Bayesian definition. Surprise is how data can affect an observer, and is measured as the difference between the posterior and prior beliefs of the observer. Hou and Zang [Hou 2007] propose a model that works independently of the input features. They remind that natural images have a 1/f decreasing log-spectrum. So, a low pass-filtering in the frequency domain to obtain the normality is applied. The difference between the normality and the normal image allow to obtain the saliency map.

3.1.2.2. Videos

Most of the existing methods for images have been extended in the case of video. Le Meur *et al.* propose in [Le Meur 2007] to add motion to spatial features. Rahtu [Rahtu 2009] add the intensity of movement to the Lab input features. Gao *et al.* [Gao 2008] replace their 2D square center-surround by a 3D (2D+ t) cubic shape. By this way, the temporal neighborhood is taken into account. Belardinelli *et al.* [Belardinelli 2009] uses a 3D (2D+t) Gabor filter banks. Bruce and Tsotsos [Bruce 2009] add spatio-temporal patches to the spatial ones for the learning of their ICA (Independent Component Analysis) model. Seo and Milanfar [Seo 2009] add the time dimension to the static model. Butko, Zhang *et al.* [Butko 2008] propose the SUN (Saliency Using Natural statistics) model based on Bayesian framework. Two methods are proposed, the first one calculate the features as a response of Difference of Gaussian (DoG) filters which can be considered as a plausible natural linear filters. The second one calculates the features as a response to filters using Independent Component Analysis (ICA) learned from natural images. This model works well as bottom up approach but it doesn't take into account any top down information. Frintrop [Frintrop 2006] propose also a model based on biological assumption named VOCUS (Visual Object detection with a Computational attention System). The model is based on an enhancement of Itti model with two different modes: An exploration one where no specific task is asked, a search one where a specified target is searched. In [Culibrk 2010], they integrate temporal information in input and by well using this information, their model manages camera having a not stationary movement. Mancas, Riche and Leroy [Mancas 2011] propose a model that detects abnormal motion. It is a multiscale approach using optical flow direction and speed. Those two features are spatio-temporally averaged on sliding windows. The spatio-temporal averages of speed and direction at two different scales are globally compared in the video using a rarity-based approach. Indeed the motion which is the most different in terms of speed and direction will have a higher saliency value as it is considered as "abnormal". They show that some movements can be more salient than others and the model works well for complex videos or dense crowds.

They show that some motion are more salient than others and works well for complex video or crowds. In [Riche 2012a], a comparative study of dynamic saliency models and human vision is done in the case of the video.

It can be seen that in most of the models applied for video, the temporal dimension is added as input features. This is due to the fact that the movement is important information in the human visual system. A few approaches analyze the information by using directly a 3D cube.

3.1.3. Saliency model including top down information

The top down information is information based on our knowledge of the environment and can be classified into two categories. The first one is based on knowledge of the normality and the previous events, the second one try to identify the normal position of objects. This information can be added to the bottom-up one. For example, if you are looking for a dog in a garden, you will search it on the ground and not in the sky.

The first category is "Attending unusual events" and is based on the knowledge acquired on the scene. Some experiments have shown that in the case of a picture, people first start to watch the center of it because, when the photographer takes it, the objects are in general in the center of the image. For a website, people watch first the top left part of the screen because it contains logo or

title contrary to the left part that contains menu. Mancas shows in [Mancas 2009] that top down information have more importance for website or advertisement. Information can be learned in video, especially for still camera. It is possible to extract a model of normality by accumulating the features of movement or to create cyclic models (traffic lights, traffic in subway). Most of the applications are for video surveillance and Jouneau and Carincotte [Jouneau 2011] used Hidden Markov Models to create their model of normality.

The second category is "Attending to objects and their usual position" and is based on the assumption of the kind of objects that can be found in different scenario and the position of objects.

Navalpakkam and Itti [Navalpakkam 2005] work on object recognition and extract features from the objects and learn them. New feature maps are obtained that will give a different importance to the parts in the sequence having the same combination of features. By this way, the saliency maps will be more linked to the kind of object searched. Another approach is the object location. The idea is to give more importance to parts of the image that have high likelihood to contain the object. Olivia *et al.* [Oliva 2003] learn the object location's and modify the saliency maps in function of the learning.

Judd propose a saliency model which takes into account low-levels features (bottom-up) but also high-levels features like face, people, and vehicles recognitions [Judd 2011].

3.1.4. Conclusion

As we have seen, most of the models are bottom-up due to difficulty to add top-down information which depends on the people. In the bottom-up models, different approaches have been tried, by comparing parts of the image with the neighborhoods, the entire image or by trying to define the normality. Work has also been done to add temporal information and saliency maps could be improved by using different cameras views, or depth information. As it is difficult to have metrics and references especially for video, comparisons of the different methods are not easy to find. Work on modeling the attention is not necessary linked to the object and others approaches perform well to define the objects but do not give any priority between them. These two points can be problematic due to the fact that it is important for decision-making to preserve the entire objects and to know which one is more important to preserve. This point should be solved. After having differentiated the important parts of the rest of the video, it is necessary to see how the video can be encoded.

3.2.Traditional encoders

3.2.1. Standards and their applications

The need to transmit images and video, regardless of place or circumstances, is growing. To respond to new applications, two great families of standards were created. ISO developed the MPEG family of standards that are rather applicable to television and HDTV. In parallel, ITU (International Telecommunication Union) proposed H.26x standards that are mainly used for videophony and mobiles phones. These standards were established to address applications for the entertainment (for example: digital television and multimedia on internet). Table 1 shows the main video and image coders and their applications.

Products	Coder	Date of the creation	content
DVD, digital television	MPEG-2	1994	Video
CD / DVD	MPEG-2	1999	Video
Blu-Ray	MPEG 2 – MPEG 4 Part 10 AVC/H264	Mai 2003	Video
Streaming on internet, Multimedia on mobile phone, video conference, video on mobile phone,	MPEG-4 ASP	1999	Video
Digital cinema, medical and outer space images	Motion JPEG2000	2004	Video
Images for entertainment, digital camera	JPEG	1986	Image
Graphical elements (logos, diagram)	GIF	1987	Image
Medical images	JPEG 2000	2002	Image

Table 1 Existing coders and their applications

We can underline that the entertainment applications have to maintain a good global quality. For each new physical support, the data storage is increased and a new video coding standard is associated. This is due to the fact that after the deployment of an entertainment standard and it associated physical support, it is not possible to change it. The increasing data storage of the support and the increasing sizes of the video push the encoders to be always more efficient at a high bitrate. Some coders are used in professional application (medicine, outer space images) but in these cases, the global quality is generally maintained at high levels. At the opposite, some works has been done to do video compression at a low bitrate for video surveillance by example. First, the principle of encoding is presented and then the traditional video coding will be review.

3.2.2. Review of encoding

3.2.2.1. Video format

Before the compression, the input video is in a raw format called YUV. The human visual system being more sensitive to variations of luminance than variations of chrominance, most encoders use YUV space, for which Y is the luminance component and UV are two chrominance components. The latter are subsampled in various ways and presented in Table 2.

Appellation (YUV)	Luminance (Y)	Chrominance (U et V)
4 :0 :0	Full horizontal and vertical resolution	No chrominance
4 :1 :1	Full horizontal and vertical resolution	¼ horizontal resolution, full vertical resolution
4 :2 :0	Full horizontal and vertical resolution	½ horizontal resolution, ½ vertical resolution
4 :2 :2	Full horizontal and vertical resolution	½ horizontal resolution, full vertical resolution
4 :4 :4	Full horizontal and vertical resolution	Full horizontal and vertical resolution

Table 2 Table of YUV subsampling

Two representations are particularly used in coding: YUV 4 :2 :0 and YUV 4 :4 :4. For the first one, luminance is at full resolution and chrominance are approximated by two, both vertically and horizontally, which gives us a quarter of the resolution and greatly reduces the quantity of data to transmit. As for the second one, both luminance and chrominance are in full resolution that allows to have a source without any loss of sample. After that, intra and inter coding can be applied.

3.2.2.2. Intra frame

Intra Coding based on spatial correlation introduces the notion of I frames and relies on the fact that one pixel or group of pixels can be predicted by adjacent pixels. Intra encoding based on spatial correlation is done without any reference to past and future frames. The image is broken down in 8x8 or 4x4 pixels in function of the coders and, for each block, a discrete cosine transform (DCT) is applied. This transformation allows us to jump from the spatiotemporal domain to the frequency domain, where quantification can be applied. During this quantification, high frequencies are deleted since they are barely perceived by the human eye and can represent an important quantity of data. Generally speaking, quantification allows to decrease the quantity of data by cutting back the set of possible values.

In a nutshell, intra images are fixed and independent of other types of images, and each Group Of Pictures (GOP) starts with one of them.

3.2.2.3. Inter frame

Inter-Frame coding depends on previous and following images. If an image is lost during transmission, it is necessary to wait for the next image that is Intra-Frame. This is where the notion “Group of Pictures” comes in. It is a set of grouped frames periodically repeated until the end of the encoded video sequence. A GOP defines the order in which are placed the images to be encoded (intra) and the ones to have a predictive coding (inter).

Inter encoding introduces P and B frames and relies on the fact that pixels do not change much in a short period of time. Motion prediction is a powerful mean to diminish temporal redundancy between images and is executed from images of reference. This concept is based on the estimation of motion between images. If all elements of the scene make relatively simple moves, motion between two images can be described with a limited number of descriptors. These descriptors, called motion vectors, are encoded and transmitted.

P frame (predictive frame) contains data on differences resulting from a compensated prediction of motion with the I frame or the previous P frame. They are also used as frames of reference.

B frame (bidirectional predictive frame), contains data on differences with past and future I or P frame that are within the GOP. In order to limit the propagation of prediction errors, B frames are usually not used as frames of reference.

3.2.2.4. GOP

Figure 6 illustrates I, P and B frames with their predictions and the GOP notion. There two kinds of GOP, the open one and the closed one. Closed GOPs cannot contain any frame that refers to a frame in the previous or next GOP. In contrast, open GOPs begin with one or more B-frames that reference the last P-frame of the previous GOP. Figure 6 illustrates a closed GOP. I_1 frame is the first image of the GOP and the reference for the P_1 frame, which itself is the reference for P_2. B_1 is predicted from the past I_1 frame and the future frame of P_1. As for B_2, P_1 is the reference of past and P_2 of the future. I_1, B_1, P_1, B_2 and P_2 form a first GOP. The following GOP starts with I_2.

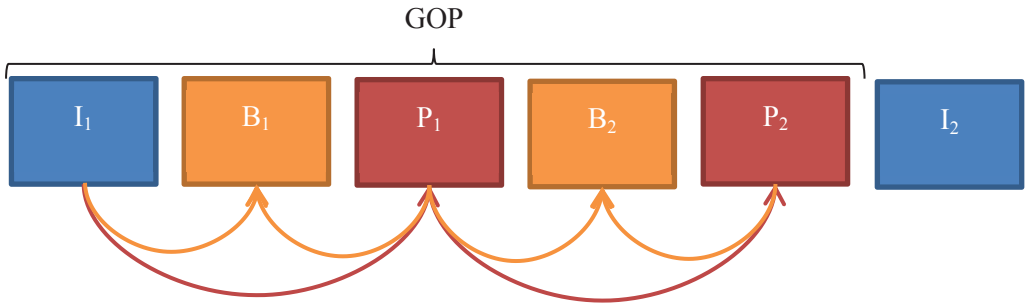


Figure 6 I, P and B frames with their predictions and notion of GOP

3.2.3. Traditional encoding methods

3.2.3.1. Hybrid predictive encoders

Hybrid predictive encoders can be defined as a coder combining a Discrete Cosine Transform (DCT) and motion compensation.

The H.261 has been the first standard for video coding and MPEG-1 part-2 has succeeded. Each image is represented in blocks and this encoder manages video at a resolution of 352 x 240 pixels at 30 frames per second (fps) in NTSC or video at a resolution of 352 x 288 pixels at 25 fps in PAL/SECAM. This encoder allows to obtain bitrate at 1.2 Mbit/s and can be used for coding on a CD. This encoder introduces the Intra coded frames (I-frames), the Predictive frames (P-frames) and the Bidirectional predictive frames (B-frames).

The MPEG-2 part 2 video coding standard (also known as ITU-T H.262) [MPEG2 1994], is a standard similar to MPEG-1 that can also handle interlaced video. Interlaced video is a way to reduce the amount of data by separating each picture into two fields: “the bottom field” contains the even numbered horizontal lines and the “top field” the odd ones. At the reception, the two fields are displayed alternatively with the lines of one filed interleaving between the lines of the previous lines. MPEG-2 part 2 specifies as MPEG-1 part 2 that the frames can be compressed with I, P and B frames. It can handle video at a resolution of 720 x 576 pixels at 30 fps for a maximum bitrate of 15 Mbit/s.

The notion of profile and level has been introduced. Profiles impose some bounds to the full syntax and define which tools or functionalities may be used and the level gives the bounds to the image size. Table 3 presents the different combination of levels and profiles available in MPEG2 Part 2.

Level\Profile	Simple	Main	SNR	Spatial	High
Low		X	X		
Main	X	X	X		X
High-1440		X		X	X
High		X			X

Table 3 Profiles and levels for MPEG-2 part 2

The MPEG-4 Visual (MPEG-4 part 2) standard [MPEG4 2001] is based as MPEG-1 and MPEG-2 on DCT. To address various applications ranging from low resolution (surveillance cameras) to high definition, MPEG-4 part 2 has provided various profiles and levels. For example, Simple profile is useful for low bitrate application. The quarter pixel motion compensation feature of Advanced Simple Profile (ASP) is one of the innovations available in MPEG-4 part 2. Another one is the video shape coding capability.

The H.264/MPEG-4 part 10 or AVC (Advanced Video Coding) [JVT 2003] has proposed number of new features that allow it to compress video much more effectively and to provide more flexibility for application to a wide variety of network environments. It proposed an Inter-frame prediction including the use of previously encoded pictures as references in a much more flexible way than in past standards, of variable block-size motion compensation (VBSMC) with block sizes as large as 16×16 and as small as 4×4, enabling precise segmentation of moving regions, of multiple motion vectors per macroblock (one or two per partition) with a maximum of 32 , of any macroblock type in B-frames, of quarter-pixel precision for motion compensation, of weighted prediction allowing an encoder to specify the use of a scaling and offset when performing motion compensation. But also a spatial prediction from the edges of neighboring blocks for intra coding, rather than the "DC"-only prediction found in MPEG-2 Part 2 and the transform coefficient prediction found in MPEG-4 Part 2. Lossless macroblock coding, flexible interlaced-scan video coding, a quantization design including a logarithmic step size control for easier bit rate management by encoders and simplified inverse-quantization scaling and a frequency-customized quantization scaling matrices selected by the encoder for perceptual-based quantization optimization are presented. An in-loop deblocking filter that helps prevent the blocking artifacts common to other DCT-based image compression techniques has been included and lead to better visual results. An entropy coding is included and contains a Context-adaptive binary arithmetic coding (CABAC) which is an algorithm to losslessly compress syntax elements in the video stream knowing the probabilities of syntax elements in a given context. In the case of lower-complexity is needed, a Context-adaptive variable-length coding (CAVLC) can be used. All these improvements, when there are well used together can give 50% bitrate savings for equivalent perceptual quality relative to the performance of the older standards.

The new standard H.265/HEVC [Han 2012] is the successor of H.264/AVC. This standard wills double the data compression ratio for the same level of quality. It can support resolutions up to 8192 x 4320 and frame rates around 150 frames/seconds. The macroblock from the previous standards has been replaced here by the Coding Tree Unit (CTU) that can be bigger than 64x64 pixels and can better sub-partition the picture into variable sized structures. The frame is divided into CTUs which are divided

for each component into coding tree blocks (CTBs).A CTB can be 64×64, 32×32, or 16×16 with a larger pixel block size usually increasing the coding efficiency. CTBs are then divided into one or more coding units (CUs) following a subdivision in quadtree. CUs are then divided into prediction units (PUs) of either Intra-frame or Inter-frame prediction type which can vary in size from 64×64 to 4×4. To limit worst-case memory bandwidth when applying motion compensation in the decoding process, prediction units coded using Inter-frame prediction are restricted to a minimum size of 8×4 or 4×8 if they are predicted from a one frame (P-frame) or 8×8 if they are predicted from two references (B-frame).The prediction residual is then coded using transform units (TUs) which contain coefficients for spatial block transform and quantization.

Predictive coders (MPEG-x) have been improved with the years and are based on metrics that measure the quality of the global image or video (PSNR-type metric) without taking into account the semantic aspect. The video at low bitrates gives block artifacts than can be mitigated by deblocking filters but the salient content can be lost.

3.2.3.2. Wavelet based encoders

Motion JPEG 2000 [JPEG2000] is the part 3 of JPEG 2000 images compression standard. The principle is to encode independently each image with the JPEG 2000 standard. The advantage of this standard is at very high bitrate because it can be used as a lossless encoding process. As there is no temporal dependence between the images, it is very useful for video editing. The JPEG 2000 standard increases the performance of image encoding compare to JPEG with a superior compression performance attributed to the use of the wavelet transform and a more sophisticated entropy encoding scheme. JPEG 2000 decomposes the images into a multiple resolution representation and a progressive decoding. The progressive decoding means that after a small part of the while file has been received, the viewer can see a low quality image. JPEG 2000 proposes also a lossless or a lossy compression schemes. The JPEG 2000 code streams are region of interest oriented and can support different quality in different parts of the same image.

3.2.3.3. Conclusion and limitation

Higher resolution images or videos leading to a more important bitrates have leaded to improve the standard. The encoders can be separated in two families, the hybrid predictive encoders based on DCT and motion compensation and the wavelets based encoders based on a wavelet transform, multiple resolution representation and progressing encoding. But at low bitrate, artifacts (blurry areas for Motion JPEG 2000 and block effects for the predictive models) appear on the images. Some encoders have the possibility to allocate more bits to a specific Region Of Interest (ROI) to preserve them but it is necessary to define them. When the bitrates decrease too much, the artifacts become too important and lead to a non-interpretable video. As in our applications, we allow to disturb the background in order to preserve the objects of interest, an overview of different perceptual coding methods have been tested. These methods give results more linked to the human perception.

3.2.4. Perceptual coding methods

3.2.4.1. Introduction

Traditional coding approaches work on the global video without adapting the compression to the visual salient parts and search to maximize the Mean Squared Error (MSE) that does not correspond to the best psychovisual result. Methods of perceptual coding have been developed to better take into account the human perception into the compression. An overview of perceptual coding for

images is first presented. Then, some references presenting an overview of perceptual video coders are described. Some specific perceptual video coders are described with their advantages and limitations. To conclude, we will highlight the limitations of these approaches in our scenarios and see the difference with the proposed approach.

3.2.4.2. Perceptual coding for images

In [Ran 1995a] and [Ran 1995b], Ran and Farvardin note that strong edges have a higher perceptual importance than weaker edges (textures) and that smooth areas influence our perception. Decomposition of the image into three components is proposed, a structure component that represents the strong edges, a smooth component that represent the background slow-intensity variations and a texture component containing the textures. The structure component is perceptually important and the intensity and geometric information are encoded. The two others components are encoded with entropy-coded adaptive DCT or subband. One of the limitation that can be highlighted is they use low levels information based only on frequency information to define what is important.

In [Bastani 2010], astani *et al.* propose an algorithm for image compression based on inpainting method. They first identify the parts of the image that can be easily recovered with inpainting and removed them at the encoder side. By this way, the spatial redundancy is reduced. The essential data to recover the entire image is encoded and send to the decoder. At the decoder side, an inpainting method is used to recover the suppressed parts. The inpainting method used is the Partial Differential Equation (PDE).The scheme leads to a very high compression ratio of 1:40 (0.2bpp) compare to JPEG for an acceptable quality. Their approach works well for a high compression and inputs with simple structures and low textures because the JPEG algorithm creates lots of block artifacts. The problem of their approach is that it works only for specific and unnatural images like cartoons.

3.2.4.3. Overview of perceptual video coding

Discussion on different perceptual video coding approaches can be found in [Wu 2005], [Chen.Z 2010], [Ndjiki-Nya 2012] and [Mancas 2012].

In [Wu 2005], Wu and Rao present in their book a review of the basics of compression, Human Visual System (HVS) modeling and coding artifacts in function of the well-known techniques. They follow by describing the subjective and objective methods and metrics, the testing procedures and international standards regarding image quality. Some practical applications such as video coder based on the Human Visual System, restoration or error correction are finally presented.

In [Chen.Z 2010], Z. Chen *et al* explain that the incorporation of the human perception in video coding system to enhance the perceptual quality become a more and more important subject. The limited understanding and high complexity of computational models of Human Visual System let this topic a challenging one. Their presentation is concentrated on three areas. First the visual attention and sensitivity modeling is treated by seeing the bottom-up and top-down attention modeling, the contrast sensitivity functions and the masking effects. Then, the perceptual quality optimization for constrained video coding is seen and finally an overview of the impact of the human perception on new different applications like High Dynamic Range (HDR) video or 3D video is presented.

In [Ndjiki-Nya 2012], Ndjiki-Nya *et al.* propose a review of the main perception –oriented video coding based on image analysis and completion. The relevance, limitations and challenges of these

coders for future codec designs is brought forward. It is also explained that work on the evaluation has to be done to obtain a new rate-quality metric.

In [Mancas 2012], Mancas *et al.* present a review of the human attention modeling and the applications for data reduction. After a presentation of the attention modeling and the saliency maps, they present works realized for perceptual coding, but also the application of attention modeling for perceptual spatial resizing.

3.2.4.4. Some specific cases of perceptual video coding

Perceptual video coding is a coding approach based on the Human Visual System and trying to give more detailed, bitrate, to the important parts. The using of a coder and attention modeling is needed. Three main approaches can be considered: The interactive one, the indirect and the direct one.

The “interactive approach” requires eyes tracking device and are consequently not common at all. The idea is to follow where the user is watching and allocate more bitrate to this region. Other problems are that it works only if there is one viewer, it is dependent of the viewing distance, and it changes also with the eye tracking system. The idea to automate this system without eye tracking device is very challenging. The use of saliency maps is necessary, and some problems may appear when, for example, no salient objects are in the scene, people will watch in all the video. Two approaches are possible. The first one is the indirect approach and will modify the video before being encoded. In this approach the coder is not modified. The second approach use direct approach and modify directly the coders.

In the “indirect approaches”, the idea is to modify the video in input in order not to modify the coder.

Itti propose in [Itti 2004] to use their model [Itti 1998] and apply a smooth filter in all the non-salient regions. This allows to have a higher spatial correlation, a better prediction and consequently to reduce the bitrate of the video. This allows to reduce by 50% the number of bit needed with MPEG-1 and MPEG-4 encoders.

Another approach proposed by Tsapatsoulis *et al.* [Tsapatsoulis 2007] combine bottom up and top down information in a wavelet decomposition to obtain a multiscale analysis. A bitrate saving of 10.4 to 28.3 % with a MPEG-4 coder is obtained.

Mancas apply in [Mancas 2007b] their saliency model in image compression. An anisotropic filtering is applied on non-salient regions and allows to decrease twice the number of bit compare to the JPEG standards. Some approach use resizing before the encoding to reduce the bitrate and obtain more flexibility on the spatial dimension of the image or video. These approaches will be more detailed in the section 3.3.5 .

In the direct approach, the coder is directly modified to reduce the quantity of information to encode. These approaches can also use image synthesis.

Li *et al.* [Li 2011] uses a saliency map to generate a guidance map that will modify the quantification parameter of the coder. By this way, more bitrate will be allocated for the salient regions. It is

underlined that some studies should be done to measure the influence of the artifacts in the non-salient areas that can become salient if the artifacts are too disturbing.

Gupta and Chaudhury [Gupta 2011] improve the model of Li *et al.* [Li 2011] and propose a learning-based feature integration algorithm incorporating visual saliency propagation that decreases the complexity of the method.

Hou and Zhang [Hou 2007] and Guo and Zhang [Guo 2010] propose approaches based on the spectrum of the images: the Spectral Residual for Hou and Zhang and the Phase spectrum of Quaternion Fourier Transform for Guo and Zhang. In the Guo approach, the object in the spectrum domain is identified and some frequencies in the background are suppressed. A bitrate saving between 32.6 and 38% compare to the traditional H.264 /AVC is reported. These methods have the advantage to be less computational intensive but they are also less linked to the Human Visual System. This approach does not work when the salient object is too important because only the boundaries will be detected and also when the background is more textured than the salient objects.

In [Chen.H 2010], H. Chen *et al.* notice that temporal prediction does not work well for video sequences with nonlinear motion and global illumination change between the frames. They propose a new algorithm for dynamic texture extrapolation using H.264/AVC encoding and decoding system. They use as virtual reference frames some synthetized frames that are built with a dynamic texture synthesis. Their evaluation was for a range of QP = {22-37} and IPPP coding. The idea is here to use dynamic texture synthesis to improve some parts of the video sequences. The perceptual results are improved for the entire video sequences.

In [Bosch 2007], Bosch integrates several spatial texture tools into a texture based video coding scheme. Different texture techniques and segmentation strategies to detect texture regions in video sequences has been tested. These textures are analyzed using temporal motion techniques and are labeled as skipped areas that are not encoded. After the decoding process, frame reconstruction is performed by inserting the skipped texture areas into the decoded frames. Some side information like the texture masks, motion parameters which ensure the temporal consistency at the decoder have to be send with the modified video sequence. In terms of data rate savings, it is shown that a combination of the Gray Level Co-occurrence Matrix (GLCM) to describe the textures and a K-means algorithm to classify them performed the best. It is shown that in average on all the sequences tested, when the quantization parameter is larger than 36, the side information become an overcost.

In [Bosch 2011], the coding efficiency of the texture based approaches relative to fast motion objects has been improved. A texture analyzer and a motion analyzer have been also tested. These methods were incorporated into a conventional video coder, e.g., H.264/AVC, where the regions modeled by both the texture analyzer and the motion analyzer were not coded in the usual manner but texture and motion model parameters were sent to the decoder as side information. Both schemes are strongly influenced by the Human Visual System (HVS). They described a set of subjective experiments to determine the acceptability of these methods in terms of visual quality. During the experiment, two sequences are compared, one with a traditional coding, and one with their approach. The subjects had no limit on making their decisions and had three options: the first sequence is better in terms of perceptual quality, the second is the better, and there is no difference between the two sequences. At a quantification parameter equal to 44, a bitrate saving around -20% for the texture based approach and around 3% for the motion based approach is reached. The

second approach gives better results because more skip block are identified but in terms of quality their subjective evaluation shows that 49% prefer H.264/AVC, only 14% prefer motion based method and 37% see no difference. Their approach based on skip block is working only on B frame. The I and P frames, opening and closing the GOP are not modified.

3.2.4.5. Conclusion

We have seen that to realize perceptual video coding, many different approaches are possible. The methods reduce the bitrate, but in general, no evaluation of the quality is done. The direct approach modifies directly the coder and work has to be done to run on different coders. As no specification about the coder has been done for this thesis and as our solution should be compatible with different coders, indirect approach has been chosen. We remind that the objective is to reduce as much as possible the bitrate in non-salient areas to increase the quality of the salient parts or to send another video. Resizing approach like seam carving has the advantage to concentrate the salient information and suppress the non-salient parts.

3.3. Seam carving

3.3.1. General approach

Seam carving is a recent approach to resize images or video sequences while preserving semantically important content [Avidan 2007]. A seam is an optimal 8-connected path of pixels on a single image from top to bottom or left to right. Formally, let *I* be an *n* x *m* image, the term *vertical seam* is defined to be the set of points

$$(1).\quad s^X = \{s_i^x\}_{i=1}^n = \{\mathrm{x}(i), i\}_{i=1}^n, s.t. \forall i, |\mathrm{x}(i) - \mathrm{x}(i-1)| \leq 1,$$

with *x* the horizontal coordinate of the point.

Note that similar to the removal of a row or column from an image, removing the pixels of a seam from an image has only a local effect: all the pixels of the image are shifted left (or up) to compensate for the missing path.

Note also that one can replace the constraint $|\mathrm{x}(i) - \mathrm{x}(i-1)| \leq 1$, with $|\mathrm{x}(i) - \mathrm{x}(i-1)| \leq k$, and get either a simple column (or row) for $k = 0$, a piecewise connected or even completely disconnected set of pixels for any value $1 \leq k \leq m$ .The cost of the seam is defined by an energy function and an energy cumulative function. Most of the works have focused on these two functions in order to improve the seam path such that important regions are preserved.

3.3.2. Energy functions for images

The objective of the energy function is to highlight the important parts of the video.

The first energy function proposed by Avidan and Shamir [Avidan 2007] is the gradient magnitude on the luminance component. To take into account the color information and improve the visual interpretation of image content, Srivastava and Biswas propose to compute seam carving in the CIE Lab color space [Srivastava 2008]. The main problem of these approaches is the gradient sensitivity to high frequencies (textures). If the salient object is smooth and the background has texture, the seam carving will suppress the salient object first. To solve this problem, Anh *et al.* use a combination of a saliency map and the magnitude of gradient [Anh 2009]. In [Domingues 2010], Domingues *et al.* propose a retargeting algorithm that is more related to human perception by exploiting an adaptive

importance map that merges several features like gradient magnitude, saliency, face, edge and straight line detection. In [Achanta 2009b] Achanta and Susstrunk apply their saliency map [Achanta 2009a] that uniformly assign saliency values to entire salient regions, rather than just edges or texture regions. This is achieved by relying on the global contrast of a pixel rather than local contrast, measured in terms of both color and intensity features. In [Hwang 2008] Hwang and Chien introduce face detection and saliency map in the energy function to improve the performance of seam carving. In [Domingues 2010] and [Hwang 2008] one of the problems is to correctly assign the coefficients to weight the different features.

As we can see, the energy map is essential in seam carving algorithm. The quality of the resizing is strongly link with the capacity of the energy map to relate the Human Visual System (HVS).To define the seams, cumulative energy functions are needed.

3.3.3. Cumulative energy function

The cumulative energy function is used to define the seam paths on the energy function.

Given an energy function *e*, we can define the cost of a seam as

(2). $E(s) = E(I_S) = \sum_{i=1}^{n} e(I(s_i))$.

We look for the optimal seam S^*that minimizes this seam cost:

(3). $S^* = min_s E(s) = min \sum_{i=}^{n} e(I(s_i))$.

These cumulative energy functions are dynamic programming algorithms and are used to find the minimum cumulative energy path and so the optimal seam. In the case of a vertical seam, the first step is to traverse the image from top to bottom and compute the cumulative minimum energy *M* for all possible connected seams for each entry *(i,j)*.

Avidan and Shamir [Avidan 2007] propose to use backward energy to find the optimal seam path, with the backward energy defined as:

(4). $M(i,j) = e(i,j) + \begin{cases} M(i-1,j-1) \\ M(i-1,j) \\ M(i-1,j+1) \end{cases}$,

e represents the energy of the image. The drawback of this method is that it does not measure the consequence of the suppression of a seam and can create some artifacts. Therefore, Rubinstein *et al.* propose in [Rubinstein 2008] to use forward energy defined as:

(5). $M(i,j) = e(i,j) + \begin{cases} M(i-1,j-1) + C_L(i,j) \\ M(i-1,j) + C_U(i,j) \\ M(i-1,j+1) + C_R(i,j) \end{cases}$,

with

(6). $C_L(i,j) = |I(i,j+1) - I(i,j-1)| + |I(i-1,j) - I(i,j-1)|$,
(7). $C_U(i,j) = |I(i,j+1) - I(i,j-1)|$,
(8). $C_R(i,j) = |I(i,j+1) - I(i,j-1)| + |I(i-1,j) - I(i,j+1)|$,

C_L, C_U, C_R represent the cost of the new pixels edges created after the removing of a seam and *e*(*i,j*) is an additional pixel based energy measure, for instance an energy function as defined in the previous section.

The advantage of this function is that it measures the energy due to newly inserted spurious edges. Indeed, artifacts may appear when a seam is removed and previously non adjacent pixels become neighbors. Recently, Frankovich and Wong propose an absolute energy which is a combination of forward energy and energy gradient [Frankovich 2011].

Tanaka *et al.* work on seam carving applied for coding [Tanaka 2010a]. They modify the cumulative energy function to take directly into account the encoding cost of a seam. The new cumulative energy function is a combination of the forward energy and seams bitrate weighted with a Lagrangian multiplier.

In our case, temporal coherence is needed. Some work has been done to improve this aspect by modifying the energy function or the cumulative energy function.

3.3.4. Temporal aspect

Rubinstein et *al.* are the first to define seam carving for video retargeting [Rubinstein 2008]. They replace the dynamic programming method of seam carving with graph cuts that are suitable for 3D volumes. To have temporal coherence, a temporal gradient is added to the gradient magnitude on the luminance component. The weight is 0.3 for gradient magnitude on the luminance component and 0.7 for the temporal one because motion artifacts are more noticeable than spatial artifacts. The seam can move from only one pixel to the left or the right following the temporal axis that can be annoying in the case of a moving object that cross the scene. They don't explain how they solve the problem of resizing continuity between GOP.

Following the idea to work in the video cube, Chen and Sen [Chen 2008] apply the seam carving not to reduce the dimension of the video but the length of it. The length of the cube is arbitrary defined. As Rubinstein, the energy map is based on a spatiotemporal gradient.

Chao et *al.* propose a solution about the problem of lack of flexibility of the seams following the temporal axis [Chao 2011]. They compute a seam at frame *t*-1 and use the black based motion estimation and Gaussian masks to predict the coarse location of the seam in the frame *t* that reduce the search range of dynamic programming and allows having seams that can move with more than 1 pixel from one frame to another.

Ishwar and Konrad propose also a solution to the problem of seams flexibility following the temporal axis [Li 2009]. They introduce the concept of ribbon and carve ribbons out by minimizing an activity-aware cost function using dynamic programming. Their model permits an adjustment of the compromise between temporal condensation ratio and anachronism of events. The problem of GOP is solved by using a sliding-window ribbon carving. By this way, they can handle streaming video.

After having seen the works done on the temporal aspect, as our goal is to do video compression based on seam carving, an overview of existing method will be done.

3.3.5. Seam carving for image and video compression approach

Seam carving has been recently applied to image/video compression.

In [Anh 2009], Anh *et al.* introduce a content-aware multi-size image compression based on seam carving. More precisely, the proposed codec encodes an image into a content-aware progressive bit stream where the seams position and value are encoded. As for the seams position, the absolute coordinate x of the first pixel of the first column is encoded and then the difference between the first pixel of the previous column and the first pixel of the current column is encoded. For the other points, the difference between the current pixel and the previous pixel of the same seam is encoded. For the color, the same concept is used by replacing the x coordinate by the absolute R, G, B values. For all the encoded information, an arithmetic coding is used. To compute the seam carving, a saliency map multiplied with the gradient is used as energy function. A Region Of Interest (ROI) bounding box is calculated with the dimensions w_0 x h_0 on the saliency map to control the reduction process and stop it when the reduced image reaches $1.5w_0$ x $1.5h_0$.

In [Deng 2011], Deng *et al.* explain that the method in [Anh 2009] is not able to adapt an image to the display smaller than the ROI size. Furthermore, since the ROI and non-ROI are encoded using different coding schemes, severe block-artifacts occur on the boundaries of the ROI and non-ROI regions. So they propose to combine the advantages of seam carving and wavelet-based coding to obtain a novel content-based spatial-scalable compression scheme. Contrary to [Anh 2009], the original image is considered as a whole and not divided into two components, while seam carving is performed in the low-frequency sub-band. The coding process is guided by the seam energy map. The SPIHT coded bit stream and the side information of the resultant seams are transmitted to the decoder. By this way, we can reconstruct the content-aware image with arbitrary aspect ratio.

Tanaka *et al.* propose an image coding scheme which incorporates seam carving [Tanaka 2010a]. An image reduced by seam carving is transmitted, along with information about the seams position. As this information is very costly, the seams positions are approximated and the seams R, G, B values are not transmitted. More precisely, the seams positions are estimated as "pillars" of length N, such that the cost of a seam is $(\log_2 3)/N$ bpp. To define the pillar length, a top down approach is applied on a modified cumulative energy function (combination of the forward energy and seams bitrate with a Lagrangian multiplier). The stopping criterion is based on the cumulative energy: if a seam has a cumulative cost superior to a threshold, the reduction process is stopped. Results are evaluated with a SSIM metric [Wang.Z 2004].

In [Vo 2010], Vo *et al.* reduce the dimension of the frame by data pruning, compress them and interpolate the frame to return at the original dimension. The data pruning is a seam carving where the seams are only straight lines. The visual distortion is less optimal than seam carving, however only the position of the line needs to be encoded. A new high level interpolation is developed. They reach a reduction of the bitrate between 5 and 8% for a H.264/AVC Quantification Parameter Intra (QPI) between 12 and 40 and QPP=QPI+1. They choose a IPPP GOP structure because the overcost of the B frames require smaller number of bits for compression and the extra bits for indicating the dropped lines become significant comparing to the bit for coding B frame.

In [Wang.T 2010], Wang *et al.* simplify seam carving principle by requiring suppression of straight lines (vertically and horizontally). Reduction of the number of frames is combined with the reduction of the spatial dimension by using this method. The energy function is a gradient computed in the same direction as seam carving. A diffusion function is used to avoid seams which are too close to one another, as it may create visual distortions. The evaluation is done with PSNR.

The method proposed in [Tanaka 2010b] improves upon [Tanaka 2010a] to limit artifacts created during interpolation. The idea is that the seams should avoid textured areas and the border of objects. If seams cross these areas or are too close to one another, artifacts may appear during reconstruction by linear interpolation. Therefore seams are required to pass through uniform areas even though these may include important regions of the image. For this purpose, a bottom-up approach is used instead of a top-down approach to define the length of an optimal pillar and to update the Lagrangian multiplier.

In [Tanaka 2010c], the authors apply the method in [Tanaka 2010a] for video reduction based on the graph cut approach of [Rubinstein 2008].The same seam is deleted for the current GOP and an 8-connectivity is allowed with the seam from the previous and the next GOP in order to avoid artifacts at the transition between GOPs. To compute the seam for the current GOP, all its frames are used but the Intra-frame is given more importance. During the evaluation, only one Quantization Parameter (QP) is used for all the frames of the two test sequences.

In [Tanaka 2011a], a piecewise linear approximation is proposed to find the optimal seam. The novelty is that the pieces of seams can have different directions and lengths.

In [Tanaka 2010a], [Tanaka 2010b], [Tanaka 2010c] and [Tanaka 2011a] the seam computation is modified by combining the forward energy with the seams encoding cost.

Tanaka *et al.* remind there are two images editing techniques, 'Seam Carving' (SC) and 'Selective Data Pruning' (SDP) clearly closely related. They can be seen has two extreme cases of the generalized SDPs (GenSDPs) proposed in [Tanaka 2011b]. The SDP is the suppression of one column or row from the image and has been applied by Võ *et al.* in [Vo 2010]. The lines are located across low frequency regions in the images. GenSDP consider both the retargeted image quality and the bitrate for side information, a suitable compromise must be taken between these two extreme cases. The following table extracts from [Tanaka 2011b] place the three approaches in terms of bitrate and retargeting quality.

Resizing method	SC	GenSDP	SDP
Bitrate for side information	High	Adaptively Changed	Low
Retargeted image quality	Excellent	Adaptively Changed	Poor

Table 4 Seam carving, selective data pruning and generalize selective data pruning performance in terms of bitrate and quality retargeting.

This GenSDP significantly reduces the required bitrates for seam path information compared with those of the original seam carving but the side information stay too important at low bitrate. Moreover their work is for the moment only defined for images.

3.3.6. Conclusion

Seam carving is a content aware resizing approach. It has been firstly design to solve problem to see images on screen with different dimensions (4:3, 16:9). This approach is based on energy function to define the important parts and a cumulative energy function to define the seam paths. To manage video, works are then been done to take into account the temporal aspect. Application of this approach for image and video compression has been proposed.

4. Proposed saliency models

4.1. Introduction

Attention is the natural capacity to selectively focus on part of the incoming stimuli, discarding less important information. It has two components, the first one is called bottom-up or exogenous and the second one is the top-down or endogenous. During the last years, lots of bottom-up models has been developed for static images but very few for video with top-down information. The idea is to develop a dynamic model that will integrate prior information to obtain a more efficient model. The model will also have an object approach, that will allow to define the salient object. Based on the work from Riche that proposes a rarity modeling for static images [Riche 2012b], [Riche 2013], temporal information has been added to obtain a dynamic model for video ST-RARE (Spatio-Temporal saliency based on RARE model) [Décombas 2013a]. In [Riche 2014] an extension of the ST-RARE model named STRAP (Spatio-Temporal Rarity-based algorithm with Priors) is proposed. The ST-RARE being a part of the STRAP model, it will not be detailed in the ST-RARE model section. The STRAP model details the mechanism of rarity and has for objective to continue to improve the temporal robustness but also to add high level priors information and to have an object approach. This has leaded to obtain a new saliency model but also manual binary mask to validate the approach.

4.2. ST-RARE model

In the ST-RARE model [Décombas 2013a], to determine the saliency of a video, the model is based on the idea that both spatial and temporal features are needed and on the assumption that locally contrasted and globally rare features are salient. This model is an extension of the static one from N. Riche [Riche 2012b], [Riche 2013] and has been extended in the STRAP model [Riche 2014]. The input features used in the ST-RARE model are both spatial (color and orientations) and temporal (motion amplitude and direction) at several scales. Following the idea that features are not necessary important information, a mechanism of rarity is applied and will be detailed in section 4.3.To be more robust to moving camera, a module that computes global motion is proposed. To be more consistent in time, the previous saliency is motion compensated and combined with the current one. This approach has been validated on the proposed data base in [Riche 2014] and is compared to the STRAP model during the evaluation.

4.3. STRAP Model

In the section, an extension of the ST-RARE model named STRAP is presented. The first model integrates temporal information to be robust in the case of video. The STRAP model continues in this idea by improving the integration of the temporal information and integrates also high priors information. This is leaded by the wish to have a saliency model that can detect objects and give a priority between them and also to use object recognition tools at the end. The proposed approach is described in Figure 7. First a temporal compensation of the movement on a sliding window is applied to handle camera displacement. In this way, neighboring frames can be analyzed together. Then spatio-temporal information is extracted from the frames during the “feature extraction” step. These features are combined together with a rarity mechanism driven by low level priors knowledge. A tracking module is used to improve the temporal stability of the saliency model. Then high level priors information like a center Gaussian or face detection can be combined with the saliency results.

A spatio-temporal segmentation is applied on the video and the result is combined with the saliency maps. This has the advantage to accurate of the results, leading to a better detection of the object.

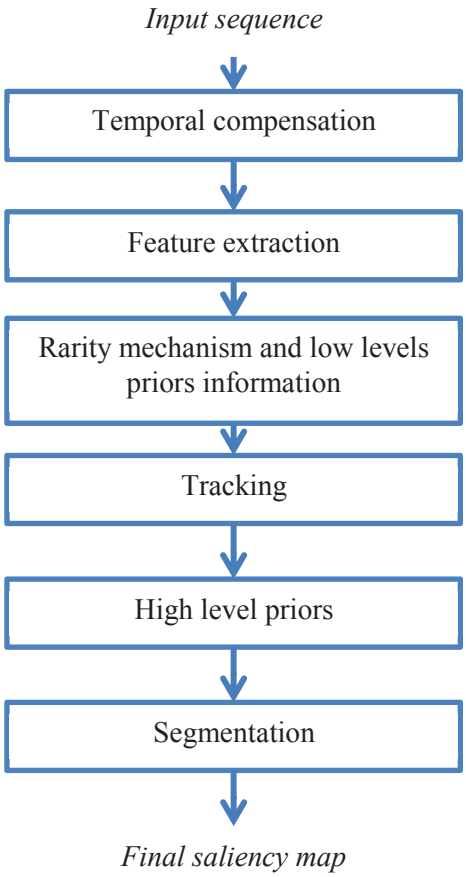


Figure 7 Proposed global approach to obtain saliency map

4.3.1. Temporal compensation

The objective of the temporal compensation is to take into account a short frame history (sliding temporal window) which stabilizes the feature extraction step. As an optimization between performance and reactivity of the algorithm a sliding window of 5 frames are used (two frames from the past, two frames from the future and the current frame). The optical flow [Chambolle 2010] is computed between the current frame and all the others and averaged into a single result. Figure 8 illustrates the approach.

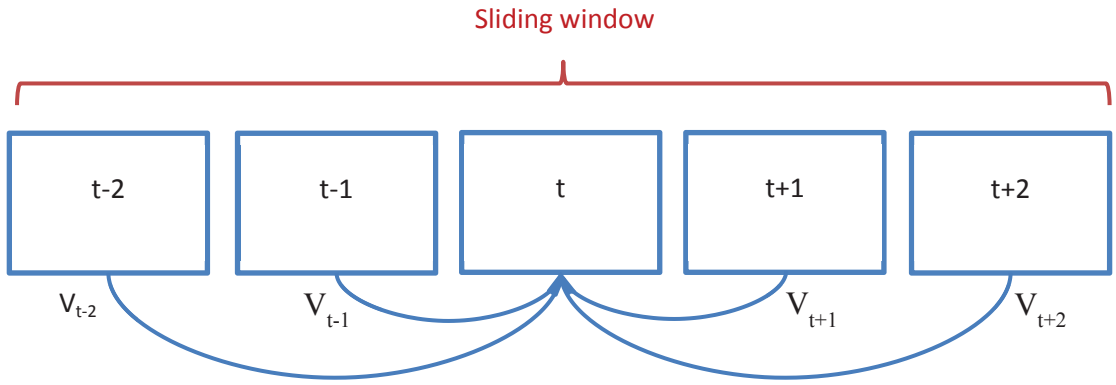


Figure 8 Temporal compensation on a sliding window.

V_{t-1} represents the optical flow vectors from the image *t*-1 to the image *t*. The neighbors' frames are estimated at the time *t* to be temporally compensated. They are then combined all together to obtain a "Combined Image".

In this way, vectors between the current frame and its neighbors and the estimated neighbors' frames can be used as input features which will be temporally more stable than the optical flow of the frame alone. The "Combined Image" is also used at the next step to extract features.

4.3.2. Features extraction

Based on the same idea than in [Décombas 2013] spatial and temporal information are extracted as illustrated in Figure 8.

Spatial features are extracted from the Combined Image. Color information is extracted using the perceptual CIE Lab color space that has the advantage to well decorrelate color information [Tkalcic 2003]. This information will be directly used in the rarity and low-level priors processing steps described in the next section.

In the same time, texture features are also extracted from each color component using a Gabor filter with 8 orientations and 3 scales. The decomposition at several scales is first combined into orientation maps. The orientation maps are finally combined together as in [Décombas 2013] and 3 textures maps are obtained, one by color component.

The temporal features are based on the list of vectors previously obtained. . To the optical flow coherence enhancement, the camera can potentially move and the background has at this moment a global motion and the others object follow their own local motion. To better identify the salient moving objects, the global motion, computed as the average horizontal and vertical movement, is subtracted to the total motion. This approach has some limitations in the case of big objects crossing the scene, but can be improved by using more complex global motion estimation [Dufaux 2000].

To better exploit the temporal features, the vectors are expressed in motion Amplitude A_t and Direction D_t defined as:

(9). $$A_t = \sqrt{\Delta x^2 + \Delta y^2}$$

(10). $$D_t = arctan2(\Delta y, \Delta x)$$

where Δx and Δy are the vector components obtained by the optical flow.

As summarized in Figure 9, six static feature maps are extracted: three low-level on the intensity ($L_{Intensity}$, $a_{Intensity}$, $b_{Intensity}$) and three medium-level on the orientation and texture information coming from the Gabor filters ($L_{Texture}$, $a_{Texture}$, $b_{Texture}$). In addition, temporal features maps are considered, expressed in terms of motion amplitude (A_{t-2}, A_{t-1}) and direction (D_{t-2}, D_{t-1}).

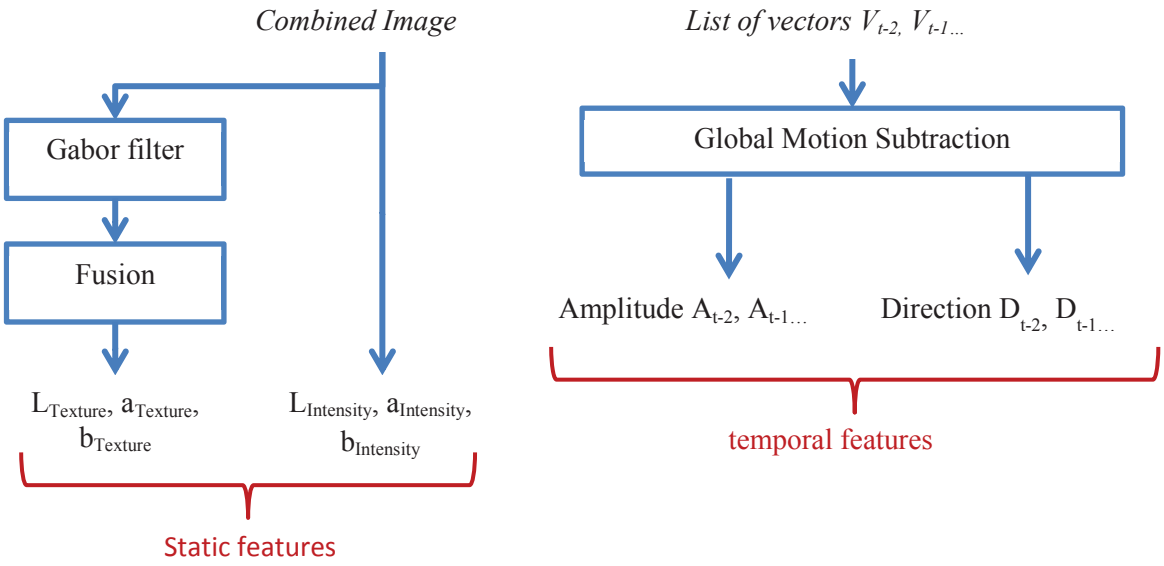


Figure 9 Static (left) and temporal (right) features extractions

4.3.3. Rarity mechanism and low levels priors information

The idea of this method is based on a set of priors from [Zhang 2013] and a rarity mechanism from [Décombas 2013]. First, two low-level prior maps are computed for spatial features maps (color and texture). A rarity mechanism is then applied on each feature map (temporal and spatial).

The first simple prior concerns frequency. The behaviour that the human visual system detects salient objects in a visual scene can be well modeled by band-pass filtering. To build this texture prior map, the authors construct a log-Gabor filter with an arbitrary bandwidth and no DC (low-pass) component [Zhang 2013]. Secondly, the transfer function of the log-Gabor filter has an extended tail at the high-frequency end, which makes it more capable to encode natural images than other common band-pass filters.

The second prior is about colours. Some studies [Zhang 2013] find from daily experiences that warm colours, such as red and yellow, are more pronounced to the human visual system than cold colours, such as green and blue. The colour feature maps are in the Cie Lab colour space. a is an opponent information on red-green information and b represents blue-yellow information. So if a pixel has a smaller (greater) a value, it would seem greenish (reddish). In the same way, if a pixel has a smaller (greater) b value, it would seem bluish (yellowish). Hence, if a pixel has a higher a or b value, it would seem “warmer”; otherwise, it would seem “colder”. Based on the aforementioned analysis, the authors perform linear mappings and build a low-level colour prior map [Zhang 2013].

Thirdly, based on the idea that a feature is not necessary salient alone, but only in a specific context, a mechanism of multi-scale rarity allows detecting both locally contrasted and globally rare regions in the image. First, a Gaussian pyramid decomposition provides feature maps at four different scales. A second step consists, for each feature, to compute the cross-scale occurrence probability of each pixel. It is obtained by the normalization of the sum of the occurrence probabilities of the pixel at all scales. Then, the self-information is used to represent the attention score for the pixel. This mechanism provides higher scores for contrasted and rare regions [Décombas 2013]. Figure 10 illustrate the rarity mechanism on a single scale.

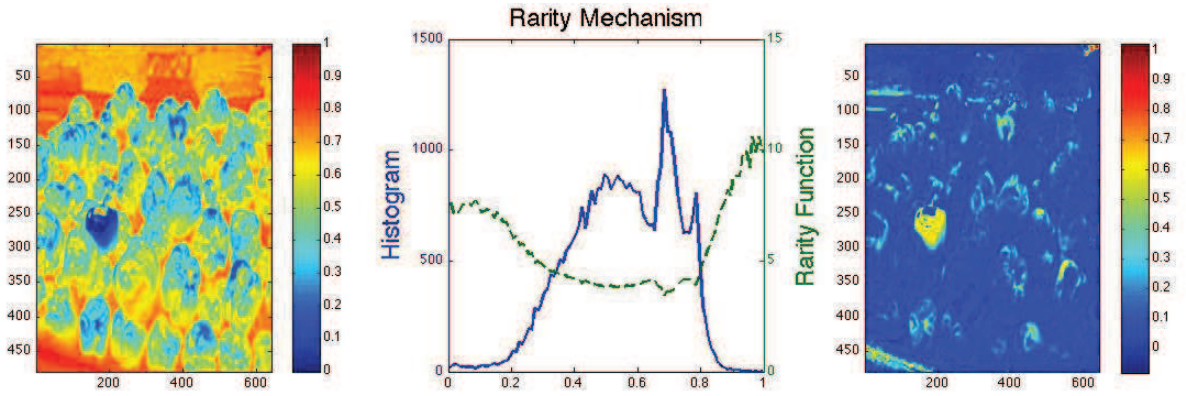


Figure 10 Illustration of the rarity mechanism on a single scale. Rarity function (green curve) is computed from a histogram (blue curve) of a feature map(left image) to obtain a rarity map (right image)

The rarity mechanism is applied first on the spatial features which are then combined with the low-level priors. Rarity spatial maps with low level priors are combined with the rarity temporal maps to obtain a saliency map.Figure 11 illustrates the approach.

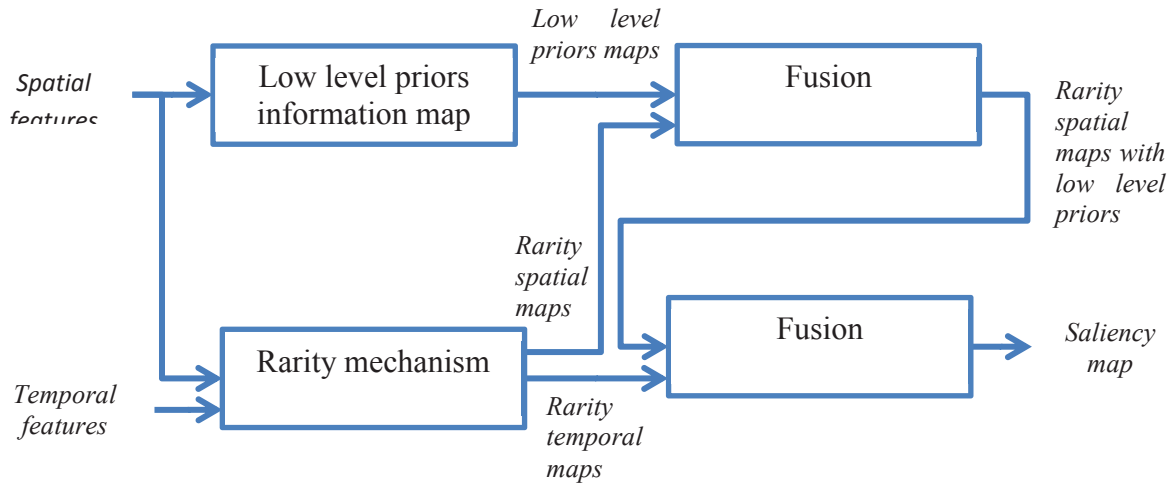


Figure 11 Rarity mechanism and low level priors information

4.3.4. Tracking

As in [Décombas 2013], a temporal tracking is used to improve the temporal coherence and robustness of the saliency results. Motion compensation of the saliency at time *t-1* is done to obtain a prediction of the saliency at time *t*. These two maps are linearly combined with a weighting factor γ empirically set to 0.3.

(11). $\text{Saliency}_{\text{final}}(t) = \gamma\ \text{Saliency}_{\text{tracking}} + (1-\gamma)\ \text{Saliency}(t)$

A high γ value will give more importance to the predicted saliency and at the opposite, the predicted saliency will be less taken into account with a low γ value. This processing results in a higher saliency value in temporally consistent regions and filters out noisy areas, improving overall robustness. The approach is illustrated in Figure 12.

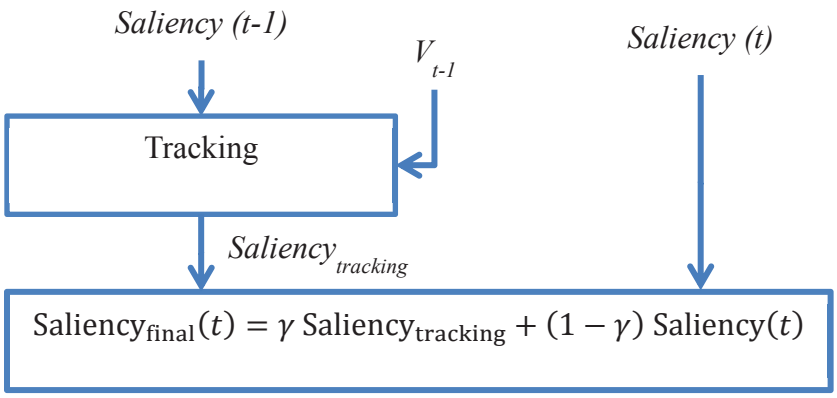


Figure 12 Saliency map tracking to obtain more robust temporal saliency map

4.3.5. High level priors

After having obtained the saliency maps, high-level priors (based on knowledge) can be added. In this paper, high-level priors are evaluated.

Previous studies [Harel 2006] have shown that the salient information is mainly located in the center of the images and the viewers have this location prior information when they are watching sequences. Human eyes gaze first in the center of the image/video which is also a good strategy to have a global perception of the scene. To model that, a centered Gaussian is used to help the model to improve its predictions. In this case, the saliency is multiplied with the Centered Gaussian.

The second model is face detection. We use the face detection algorithm from [Zhu 2012]. Following the idea in [Marat 2010], the number of detected faces and their size influence our attention. To model these observations, each detected face is weighted in function of the total number of detected faces detected. So, one face alone will have more importance that 4 faces, mechanism which is similar to the rarity approach on low-level features.

A Gaussian filter is applied on each detected face. The sigma of this filter is directly linked to the mean size of the faces detected: a big face will have more importance than a small face. In this case, the face detection is linearly combined with the saliency result..

So high-level priors maps are combined to the saliency maps as illustrated in Figure 13. In the absence of high priors information, the high priors map is simply the identity. The two models, based on center-bias and face detection, can also be combined together. After the fusion, the result is defined to well predict the human visual system without taking into account the shape of the objects.

Saliency map | High-level priors map | Saliency at high levels

Figure 13 Combination of the saliency map with high priors maps

4.3.6. Segmentations

When viewers are watching a salient object, they focus their attention on a specific point of an object. Traditional models only provide picks of saliency without taking into account the shape of the salient objects to which the pick is superimposed. To have a more object-oriented approach, color segmentation algorithm is used on a group of pictures. The model used here is an Expectation Maximization of Gaussian mixtures [Dempster 1977].

The segmentation algorithm is applied on the Cie Lab color space, separately on the Luminance and the Color information (a,b) due to the fact that the luminance is decorrelated from the two colors components that are correlated together. The algorithm is applied on a group of images to have a temporal continuity. The two segmentations maps (color and luminance) are combined together and then a spatial fusion is applied. The very small regions having an area inferior to a threshold T_1 are fused with the neighbor regions having the nearest colors. Regions with an area inferior to a threshold T_2, with $T_2>T_1$ and a color distance inferior to a threshold T_c are combined. Visual experiments has been done to fix $T_1=5$, $T_2=10$, $T_c=500$ for all the sequences. The first fusion allows suppression of very small regions that can be considered as noise or artifacts and the second step allows the fusion of small similar regions while keeping small salient regions (for example the eyes on a face). Figure 14 summarizes the combination of the high-level priors saliency map with the segmentation maps. At each region of the segmentation maps is associated the mean saliency of the region. This final result is a saliency map which highlights objects in the image and it is no more a fuzzy area of it. It is much easier to recognize the salient object when just looking to the saliency map.

Figure 14 Combination of the saliency model with the segmentation

4.4.Conclusion

In this chapter, saliency model based on rarity has been presented. The first one, called ST-RARE [Décombas 2013a] is a temporal extension of the model of N. Riche [Riche 2012b], [Riche 2013]. This ST-RARE has been extended in [Riche 2014] to obtain (2) a spatio-temporal rarity-based algorithm with priors for human fixations prediction and objects detection in videos called STRAP has been developed. We propose, (2a) a temporal compensation of the movement on a sliding windows allowing to manage both static and moving cameras and (2b) giving more robust spatial and temporal features. (2c) These features are combined together with a new model based on rarity and low level priors. (2d) High level priors information are combined to the salient models to increase the performance and (2e) a segmentation model is used to have a more object based approach. To validate the results, the raw format database with eyes tracking results from [Hadizadeh 2012] is

completed with (2f) manual binary masks to evaluate the detection of the salient objects. This database has been chosen due to the fact that it is free of artifacts. As lots of saliency models are proposed and are validated on different database, (2g) we will compare our results with different high level priors information to be consistent with the other models. 7 models are chosen for the evaluation on the 3 references (eyes tracking first view, eyes tracking second view, binary mask) with 4 different metrics. The experiments and the limitations of the approach will be detailed in chapter 7.

5. Proposed video coding based on seam carving

5.1.Introduction

The objective of traditional video coding approaches like H.264/AVC and HEVC is to minimize Mean Squared Error (MSE) but they do not explicitly consider visually salient regions. From a psycho-visual point of view, these approaches may not be optimal. For defense and security applications, with limited infrastructure and bandwidth, the video transmission is often constrained to low data rate. In this context, the transmitted information has to be the most pertinent for human understanding, at the lowest possible rate. The *overall* image quality, as generally estimated by video codecs, is not a well-suited criterion. The objective is that users can correctly interpret the content and take decisions in critical conditions by maintaining the semantic meaning of the sequences.

For this purpose, a semantic content aware video coding scheme based on seam carving is proposed. The advantage of using seam carving for video compression is that salient information is concentrated in a reduced resolution sequence and the non-informative background is suppressed, leading to a significant bitrate reduction and a better preservation of the salient parts. Since the seam carving process is not reversible, the correct position of the objects may be lost during the seam reduction. It is therefore necessary to transmit some additional information about the seams in order to properly recover the original dimension of the video sequence and the position of the salient parts. In most of the papers using seam carving for image or video compression, the cost of the seams texture is too important and synthesis algorithms are usually applied to find the background. As our objective is to concentrate on the salient information, no background texture synthesis will be applied in the proposed approach. Nevertheless, any kind of inpainting algorithms [Bertalmío 2000] and [Bertalmío 2001] or others synthesis algorithms could be used.

The different approaches proposed aim at modeling the seams to avoid a significant overhead and to well position the salient object at the decoder. The main idea of these approaches is based on the observation that the seams are more concentrated between the salient objects and can be seen as group of seams. Figure 15 illustrates this observation. It can be observed on the image in the center that the seams are concentrated in group of seams between the salient objects. On the image on the right, the seams are suppressed in the original image and a reduced one is obtained that will be encoded. It can be seen that the salient information is well-concentrated.

Figure 15 Illustration of the group of seams for ice. From left to right: original image, original image with the seams in orange, reduced image.

To represent the seams, two main approaches have been proposed. The first one is based on a encoding of some key points that will be used at the decoder to modify the cumulative energy maps.

In these approaches, the seam shapes are not totally known and saliency maps are computed at the decoder to control the seams in the unknown parts. These approaches has been published in [Décombas 2011], [Décombas 2012a] and [Décombas 2012b]. The second approach is based on a modeling of the seams at the encoder side and as the seam shapes are totally known, it does not need saliency maps at the decoder. It has been published in [Décombas 2014]. In the two approaches, some common innovations have been introduced.

5.2.Common parts of the proposed approaches

5.2.1. Global approach

Figure 16 shows the global approach. Seam carving is first used to reduce the dimension of the video sequence as much as possible, while still preserving the salient objects. Then, the reduced video is encoded with a traditional encoder like H.264/AVC and the seams are modeled and encoded with our proposed scheme. After transmission, the video sequence is reconstructed at the decoder side, in order to recover the original dimensions and to preserve the scene geometry.

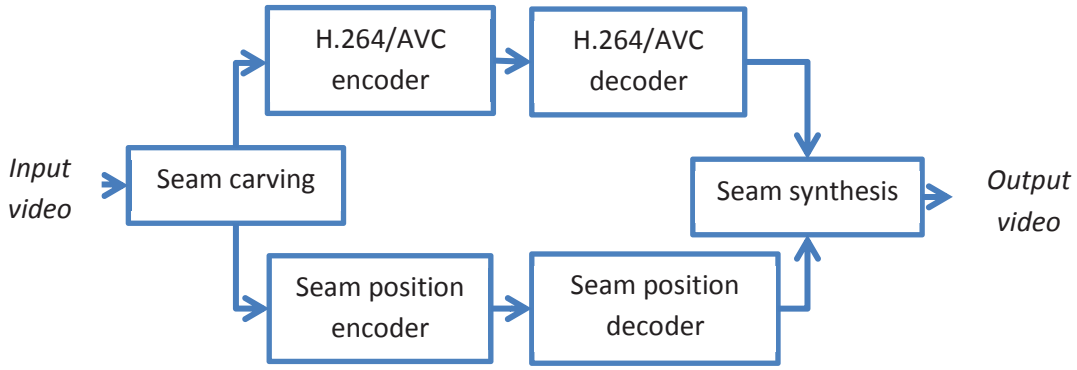


Figure 16 Architecture of the proposed semantic video coding using seam carving.

As seam carving is a process that reduces one dimension at the time, it is necessary to apply it twice. On Figure 17, this process is illustrated. First, seam carving is applied on the image to reduce it vertically, and a list of vertical seams is obtained. Next, the image is rotated by 90° clockwise and the same process of seams carving is applied. In this way, the image is horizontally reduced and a list of horizontal seams is obtained.

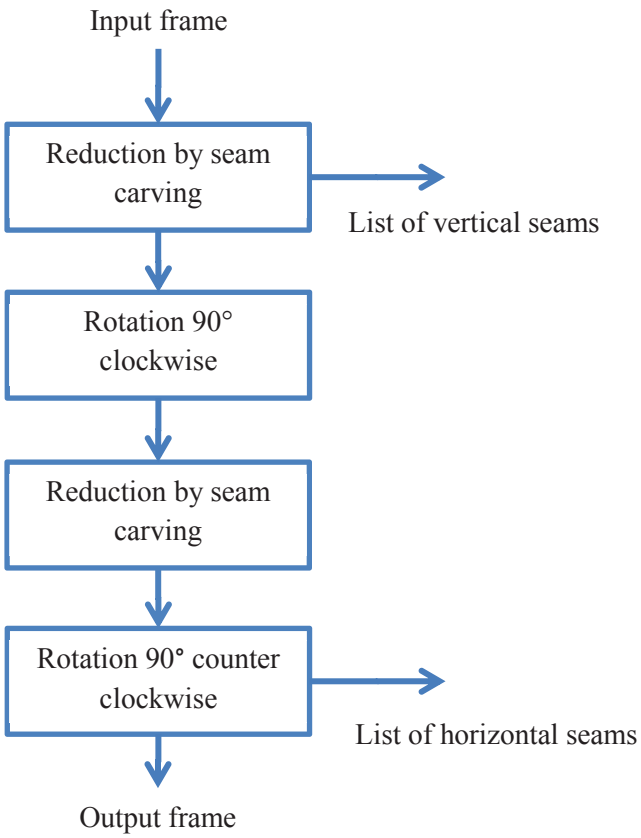


Figure 17 Vertical and horizontal seam carving reduction.

It can be pointed out that the final result will depend on the order in which the seam carving process is applied. Moreover, the result of seam carving depends on the way it is applied: the seams computed from top to bottom are different from the ones computed from bottom to top. In our case, these orders have been empirically defined, the seams are first vertically suppressed from the bottom to the top and then horizontally suppressed from the left to the right.

In Figure 18 the seam carving process is more detailed. Seam carving is applied to each frame of the GOP. For each frame, an energy function is defined based on gradient, saliency and temporal information. Post processing is applied on this energy map to better preserve salient objects. The forward cumulative energy map is then computed to define the seam to be suppressed. In parallel, the energy map is binarized to obtain a control map. This control map is used to decide when to stop seams suppression. More precisely, seam carving is iterated until reaching an object in the control map. In this way, a list of seams is obtained. Different methods have been proposed to model the seams and are more detailed in section 5.3 and section 5.4.

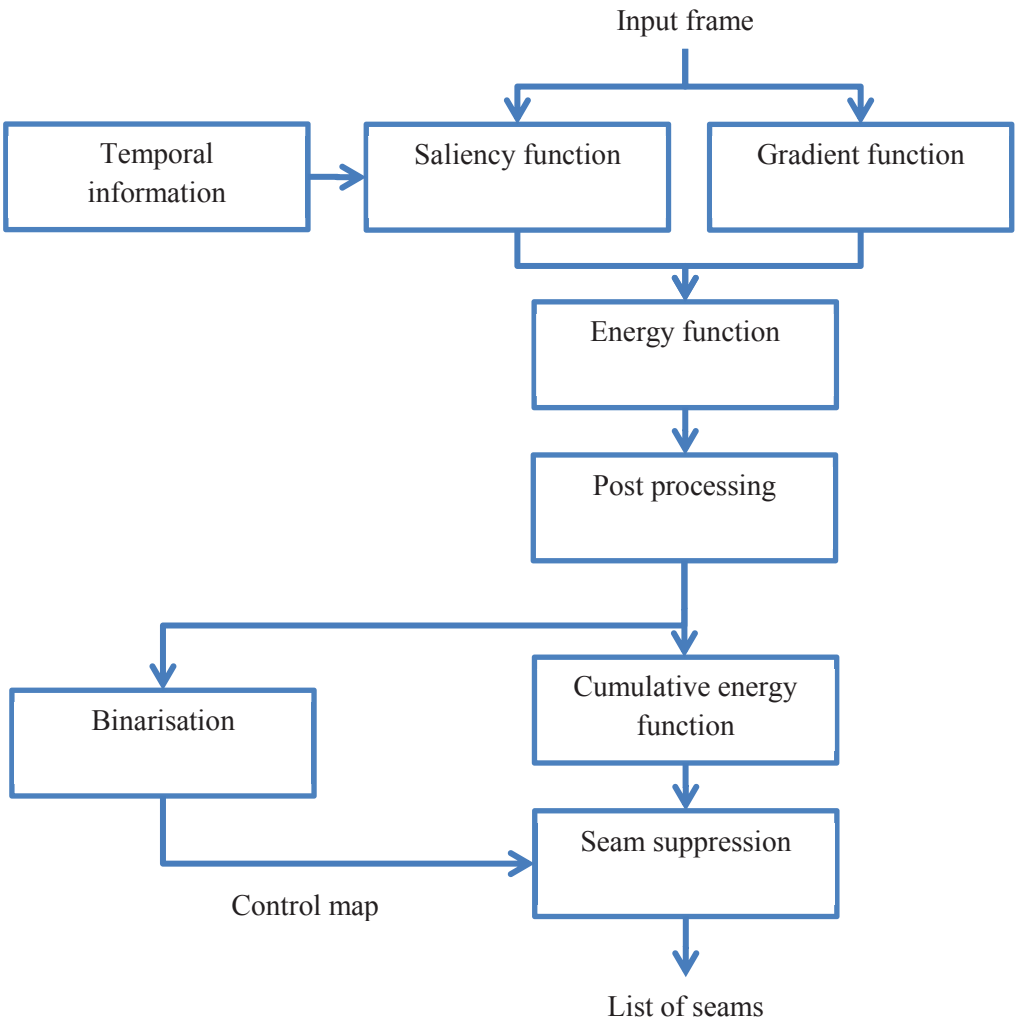


Figure 18 Overall proposed seam carving scheme.

The use of seam carving for video compression highlights two constraints during the encoding of the reduced sequence. To avoid padding during coding, the number of seams suppressed should a multiple of 16. Moreover, as seam carving changes the spatial dimension of the video, it is necessary to keep the spatial dimension of the video constant within a GOP. Figure 19 illustrates these constraints. The number of seams that can be suppressed is used to control the seam carving process. In this way, a list of seams and a reduced GOP are obtained.

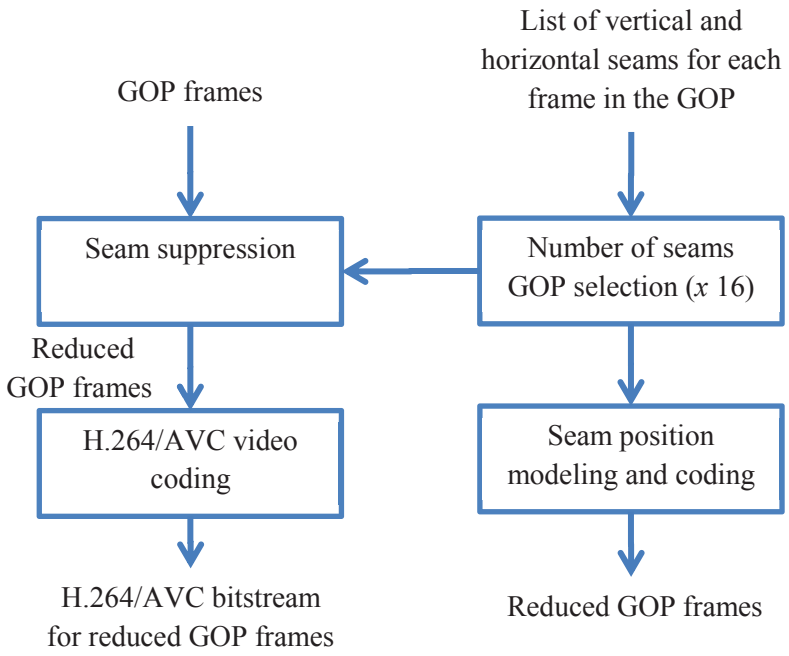


Figure 19 GOP-based seam carving process.

The coordinate of the seams are defined relative to the original image dimension. Next, all the seams are rearranged by ordering the horizontal, respectively vertical, coordinates in an increasing order. In this way, the seams are defined from left to right, respectively top to bottom, and they do not intersect. This is illustrated with three seams in Figure 20.

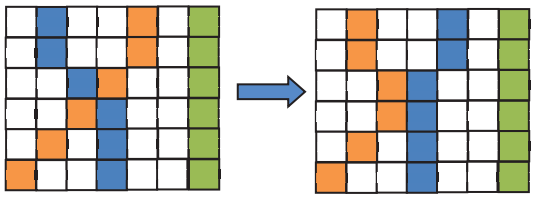


Figure 20 Seams reordering: On the left, seams before reordering, on the right, seams reordered.

To illustrate the different modeling, Figure 21 represents in different colors the disposition of 7 seams.

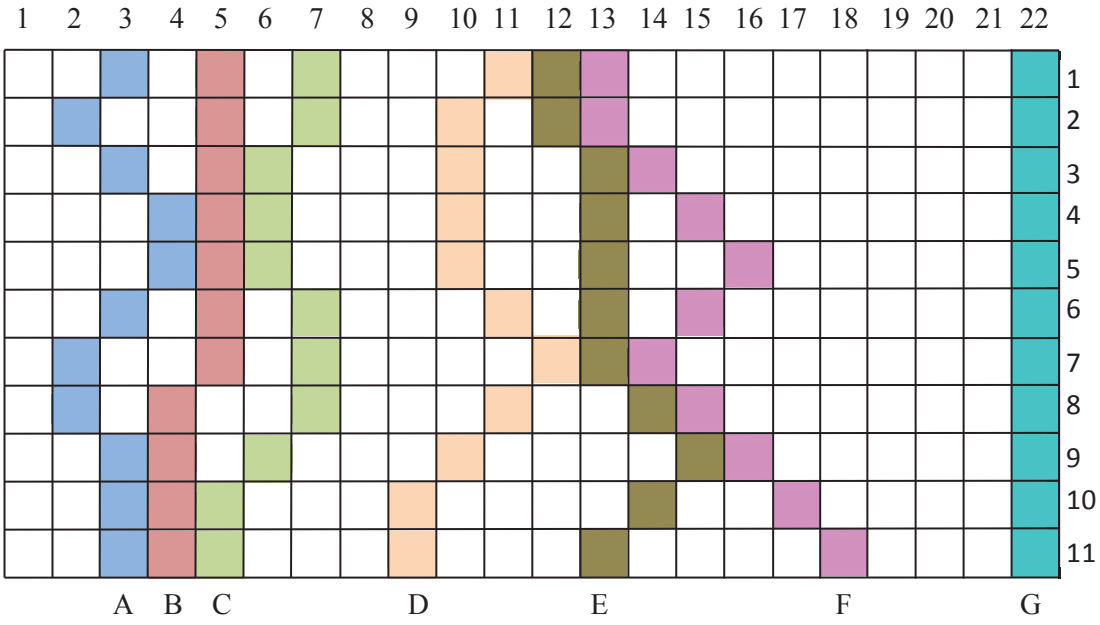


Figure 21 Diagram of seam carving with 7 different seams to explain the seam modeling

5.3.Approaches based on saliency map at the decoder

In these approaches, the size of the GOP has been arbitrary defined and is constant. In a GOP, the number of seams suppressed is constant and is defined by the frame having the smallest number of seams suppressible.

The main idea of these approaches is to identify the lines where the seams are concentrated and to encode theses points to well reposition the salient object at the decoder side. To define the important points and to encode the seams, different approaches have been tested.

5.3.1. Encoder side

5.3.1.1. Approach in [Décombas 2011]

In this approach, the idea is to identify some key lines, defined as lines where seams are concentrated. The position of the seams on these lines will be encoded to be transmitted.

First, a clustering is done for each line perpendicular to the seam carving path. We scan the position of the seams and regroup them together. In other words, for vertical seams we are using horizontal scan lines (rows) and for horizontal seams we are using vertical scan lines (columns). More precisely, given a current seam position, if the distance to the next seam position is under a threshold *T_Gpe*, we add the position to the group, otherwise we create a new group. To give an example of the distance between two neighboring pixels in the same row is zero, for example pixel at (m-1,n) and (m,n), where m is the column number and n is the row number. Then pixels at (m-1,n) and (m+1,n) are at distance of 1, i.e., there is one pixel (m,n) between them. The process is iterated until the end of the line. By this way, the seams are grouped together for each line.

The next step is the selection of key lines on a predetermined number of lines *Nb_Key_Lines*. To initialize the process, we select equidistant reference lines. Then, each line *y* oscillates in the interval $\left\{y - \frac{K_{Oscillate}}{2}; y + \frac{K_{Oscillate}}{2}\right\}$ and the line having the largest group is selected. Afterwards, we save

the position of the key line and all the group of seams position on this line. We underline that it is not judicious to keep only the position of the largest group of seams on the key lines as it may results in shifting and distorting the frame during seams synthesis.

Figure 21 illustrates our parameters in the case of a vertical seam carving for a 11x22 pixels image. If *T_Gpe*=0, the third line (*y*=3) is composed of 5 groups, three of 1 seam, two of 2 seams. If *T_Gpe*=2, there are two groups of 3 seams and one of 1 seam.

We continue our illustration by fixing *T_Gpe*=0 and a number of key line *Nb_Key_Lines*=2. The key line is initialized at the lines *y*=3 and *y*=8. The next step is to find the largest group of seams. If $K_{oscillate}$=0, the key lines selected are *y*=3 with two largest groups of 2 seams and *y*=8 one largest group of 2 seams. If $K_{oscillate}$=4, the key line selected is the line *y*=1 with the largest group of seams having 3 seams and the second key line is arbitrary defined between the lines *y*=7,10 having all one largest group of 3 seams.

For each key line, we save the position of the line *y*=1, the position of the groups of seam, *x*=3, 5, 7, 11, 22 and the number of seams inside which is 1, 1, 1, 3, 1.

Figure 22 resumes the approach in [Décombas 2011]. From a list of seams, the distance between them is computed to create group of seams in each line. The lines having the most important groups of seams are selected and the position and quantity of seams on these key lines are saved.

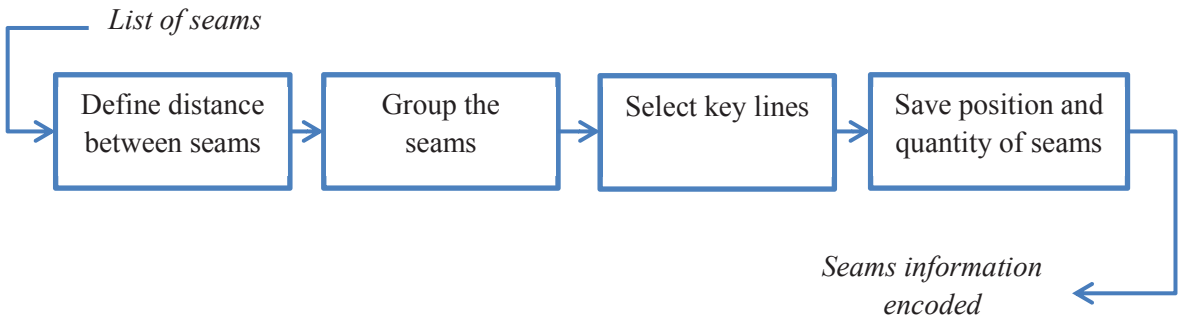


Figure 22 Selection of key points with the approach [Décombas 2011]

Figure 23 illustrates the approach on the ice sequence. It can be seen that seams are concentrated between salient object on the image on the left. The black lines show the saved position of the seams that are preserved in the right image. Between the key lines, the seams position are not preserved.

Image with original seams in orange and key lines in black

Image with approximated seams

Figure 23 Illustration of the seam carving approximation on image extracted from the Ice sequence. Left image: Original seams in orange and key lines in black. Right image: Approximated seams in orange

This approach allows to validate the concept to identify key lines for encoding but it needs many parameters and select the lines having the largest group of seams without taking into account during the selection the others seams in the line.

5.3.1.2. Approach in [Décombas 2012a]

In this approach, the groups of seams are identified by clustering, where the keys points can be in different lines. Isolated seams are also discarded. On Figure 23, the right vertical seam can be considered as an isolated seam.

The process of seam discarding is used when "the number of seams selection" from Figure 19 is applied. More specifically, the most isolated seams are removed. This idea is motivated by the fact that an isolated seam is costly to encode, can create some high frequencies in the reduced frame leading to a potential increased coding cost. In details, for each seam, the distance between the current seam and the seam at its left, respectively at its right, is computed. Next, the lowest value among the two distances is kept and associated to the current seam. Finally, the seam having the largest distance is suppressed. The process is iterated till the frame size reaches a multiple of 16 to avoid pixel padding.

About the clustering, it is used to group the seams whose coordinates will be transmitted. For this purpose, a k-median algorithm is used. The number of median seams is defined by the user. Hereafter, the number of median seams has been empirically chosen in a range between 3 and 6. Each seam is then associated to a cluster by minimizing its Euclidian distance.

The median seams are used to define the coordinates of each cluster. Given that the seam carving reinsertion process at the decoder is performed from left to right and therefore pushes the right part of the image, it is necessary to translate the median seams coordinates accordingly. In this way, after the reinsertion, the groups of seams are properly centered.

We now identify concentration areas in the perpendicular direction in order to define group of seams. The coordinates of these concentration areas have to be encoded to correctly reposition the

salient objects at the decoder. To find these areas, for each median seam, the cumulative Euclidean distance between the median seam and its associated group of seams is calculated. This is done for each point of the median seam. The concentration areas are then defined as the points with shortest distances.

To avoid keeping many points from the same concentration area, the seam is beforehand divided into subparts and the minimum cumulative Euclidean distance is searched for each subpart. The number of subparts is defined by the user. Clearly, defining more subparts results in more precise seams at the synthesis. So for each subpart and for each median seam, the concentrated area is defined as the position of the line having the minimal cumulative distance between the median seam and the associated seams.

To illustrate the approach on Figure 21, 3 median seams are searched and the seams are divided into 2 subparts. For the first median the lines with y=4 or 5 and the line with y=10 or 11 are selected. For the second, the line y=1 and y=7, for the third, there is no specification to choose a line or another one so y=3 and y=8.

5.3.1.3. Approach in [Décombas 2012b].

In this approach, we improve upon [Décombas 2011], [Décombas 2012a] on a specific point. After having defined the seams to suppress, a resized image is obtained and the seams are approximated by the approach from [Décombas 2012a]. Instead of transmitting this resized image, a new one is defined from the original image and the approximated seams. So the concept is to include the decoder part inside the encoder. By this way, the seams from the encoder parts are more similar of the seams at the decoder parts than the original seams. This closed loop process leads to less geometric deformation when the initial seams are spatially scattered in the sequence.

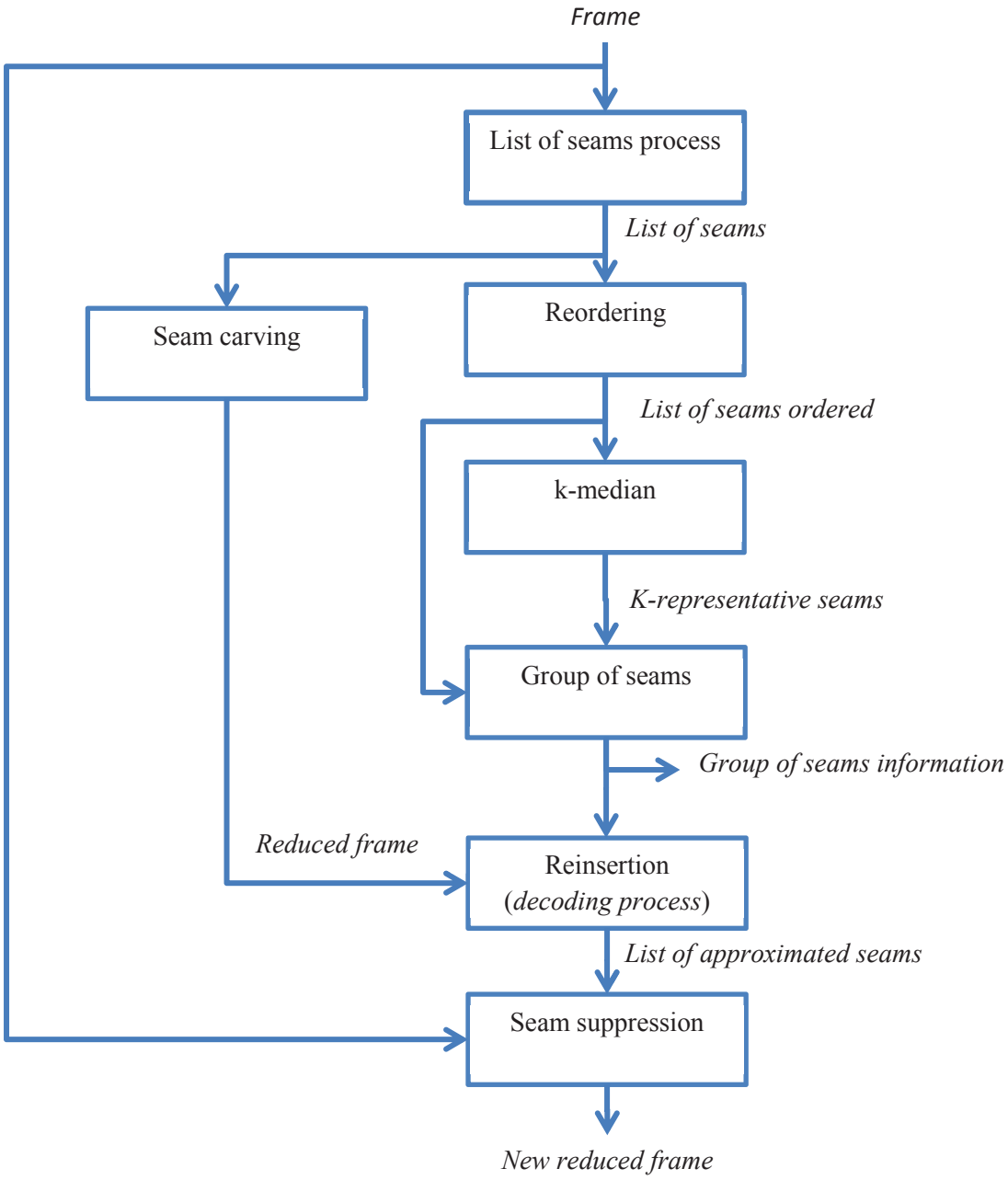


Figure 24 Closed loop approach to constraint seam carving.

Figure 24 illustrates the closed loop approach. The decoding process is simulated at the encoder side to have a list of approximated seams. This list of seams is then used to obtain a new reduced frame that will be encoded. The advantage of this method is there are less geometric deformations in the background but the reconstruction is still not totally controlled. This is due to the fact that the seams are computed on the reduced image.

Figure 25 illustrated the different approaches on an image of the container sequence. It can be seen that the different due to the approximation is more important between the left image and the center image than the center image and the right image. This will lead to less geometric distortion on the image at the decoder.

Original seams at the encoder

Seams at the decoder with the approach from [Décombas 2011], [Décombas 2012a] and at the encoder with the approach [Décombas 2012b]

Seams at the decoder with [Décombas 2012b]

Figure 25 Illustration of the seam carving approximation on image extracted from the container sequence. Left image: Original seams at the encoder. Center Image: Seams at the decoder with the approach from [Décombas 2011], [Décombas 2012a] and at the encoder with the approach [Décombas 2012b]. Right image: Approximated seams at the decoder with the approach [Décombas 2012b]

5.3.2. Decoder side

The information about the seams from the encoder is used at the decoder to reconstruct the seam path. Figure 26 illustrates the proposed seam carving process at the decoder size.

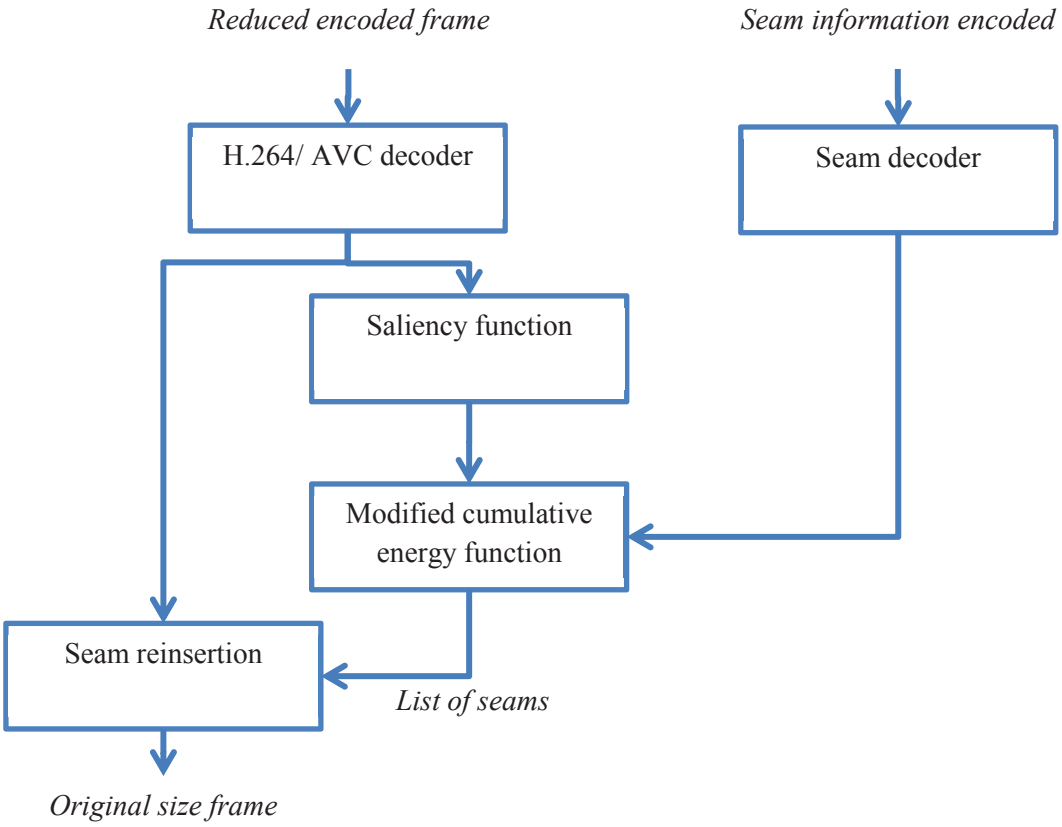


Figure 26 Proposed seam carving process at the decoder size

First, the seams and the reduced frames are decoded. From the reduce frame, a saliency map is computed and the seam paths are defined with the modified cumulative energy function. A normal cumulative energy map cannot be used because the seam carving is not a reversible process and the position of the seams at the decoder will not be the same than at the encoder. It is necessary to modify the cumulative energy map to control the seam path by using the seams information decoded. For each key lines, the cumulative energy function is modified by adding a *Saturated_value* on all the lines except at the position of the group of seams. In this way, the seam is free to find an optimal path between the key lines and in the key lines its position is totally controlled. Figure 27 illustrates the influence of modifying the cumulative energy map. The key line is saturated and the potential positions of the seam are represented in yellow. A list of seams is obtained and used to reconstruct the frame at its original size. With this approach, the salient objects are well repositioned.

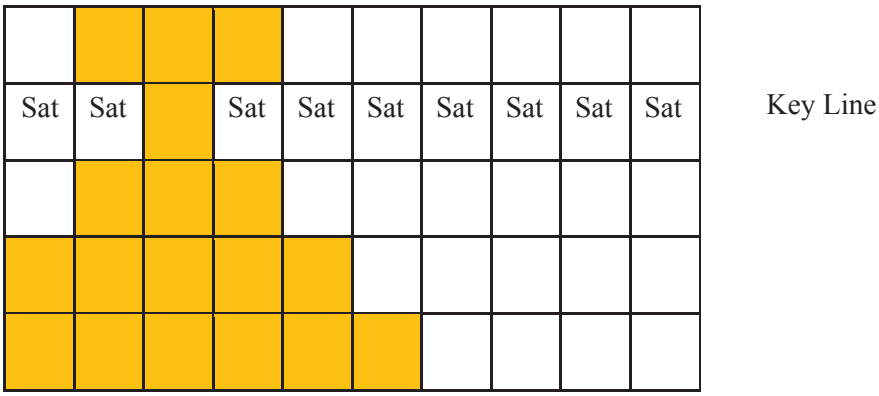


Figure 27 Modified cumulative energy map

5.3.3. Limitation of these approaches

These approaches, based on a saliency map at the decoder, perform well but have some limitations. The exact reconstruction of the seams cannot be known at the encoder. In particular, the saliency map is computed on a reduced and compressed frame, leading to some differences difficult to predict. In addition, the approximation between the key lines is efficient to define the optimal path but the final result can lead to some important geometric artifacts. For these reasons, the next proposed approach aims at totally modeling the seams at the encoder.

5.4.Approach without saliency maps at the decoder [Décombas 2014]

In this approach [Décombas 2014], the principal idea is to model the seams at the encoder side to totally control the positions of the seams. In this way, no salient maps are needed at the decoder. Figure 28 details the seam carving block from Figure 16. During the seam carving process, lists of seams are computed for each frame of the sequence to identify the variation of the number of seams in time. The sequence is then adaptively subdivided into GOPs. This is realized during the content aware GOP cutting. Then, for each GOP, a spatio-temporal seam clustering is performed, allowing to discard isolated seams and to model the groups of seams. In this way, a list of modeled seams is obtained for each frame of the sequence and they are suppressed from the original video sequence to obtain a reduced video.

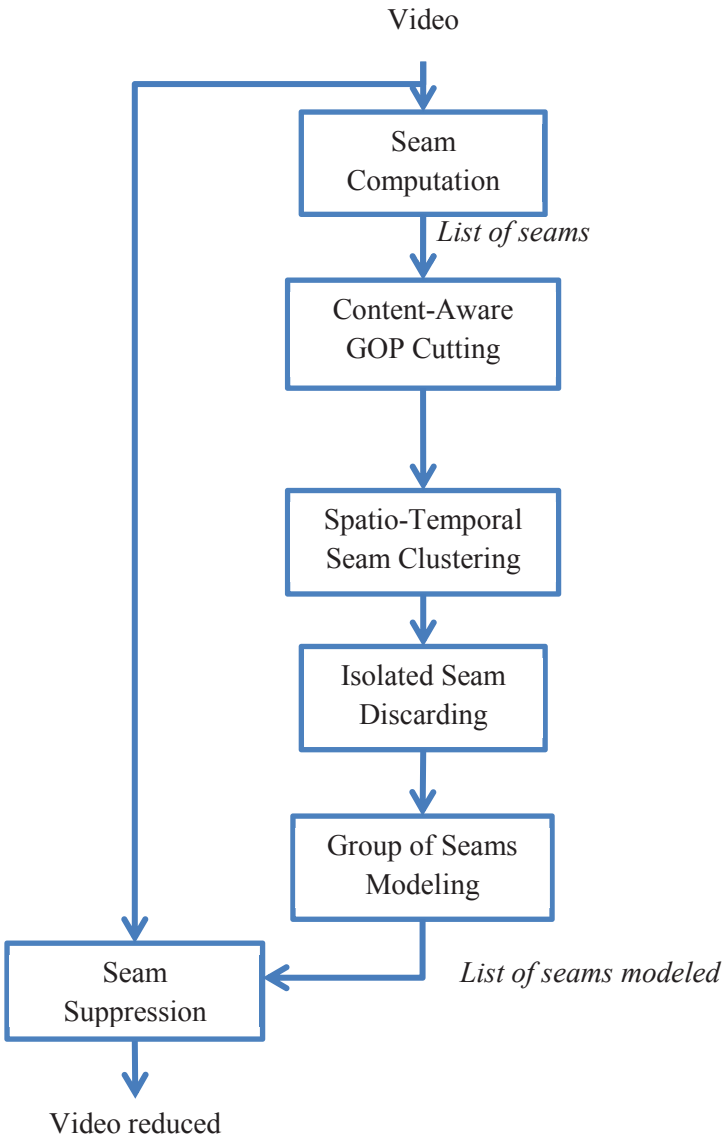


Figure 28 Overall scheme of the proposed seam carving process without saliency maps at the decoder.

The seams computation module depicted in Figure 28 is further detailed in Figure 29. An energy function is defined from the saliency model in [Riche 2014] (as described in section 4.3) for each frame. This model identifies the salient objects by finding the rarity on different maps. The most pertinent maps are combined together to obtain a unique saliency map. The model uses static (L,a,b) and dynamic (amplitude and direction) components in order to identify salient areas for static and moving scenes. A spatial median filter is then applied to delete the noise and a dilatation filter is applied to preserve the salient objects and their neighborhoods. The forward cumulative energy map is then computed to define the seam to suppress. In parallel, the energy map is binarized to obtain a control map. The binarisation is automatically defined with the Achanta approach [Achanta 2009a] that defines the threshold of binarisation as $T = 2x\ mean(Saliency)$. The control map is used to decide when the process of seam suppression is stopped. More precisely, seam carving is iterated until reaching an object in the control map. Thus, a list of seams is obtained for each frame. The process is successively applied vertically and horizontally.

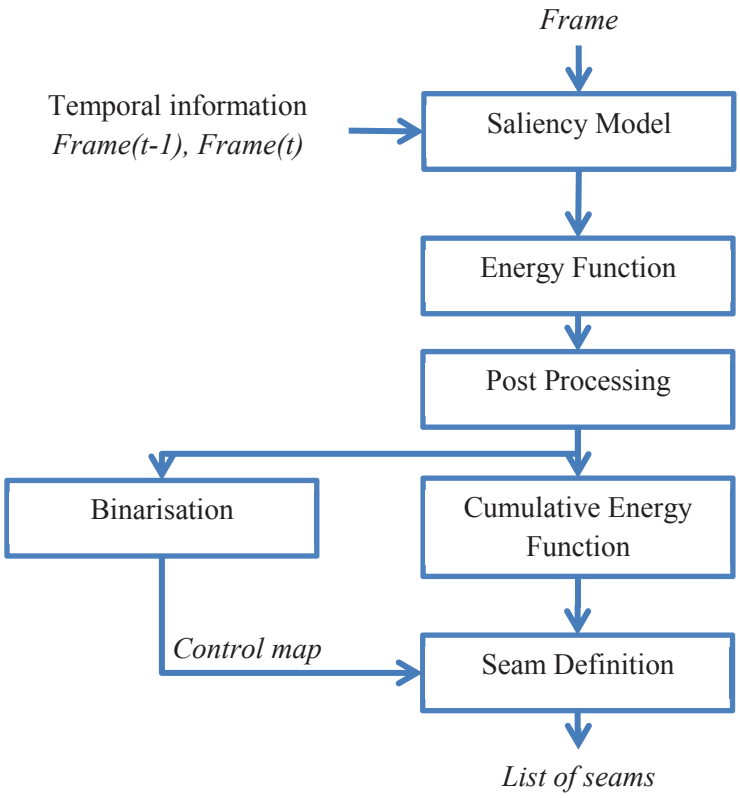


Figure 29 Overall process of seam computation.

5.4.1. Content-aware adaptive GOP cutting

To use seam carving in a video compression application, the sequence is divided into GOPs. In a GOP, in order to avoid padding, it is preferable that all the frames have the same spatial dimension, defined as a multiple of 16 pixels (corresponding to the size of a macro block in the subsequent video coding scheme). In our previous works [Décombas 2011], [Décombas 2012a] and [Décombas 2012b] (see Sec.5.3), the length of the GOP was predefined and the number of seams suppressed for each GOP was linked with the frame having the largest salient parts. That led however to a suboptimal reduction of the sequence dimensions. In this paper, we propose to adaptively split the sequence as a function of the number of seams that can be suppressed. In this way, the GOP can be further spatially reduced, without damaging the salient objects and the quantity of information to transmit is lowered. Three main cases can be identified: an object of interest is appearing into the video and the number of seams that can be suppressed is decreasing; an object is disappearing from the video and the number of seams that can be suppressed is increasing; a constant number of still or moving salient objects are present in the scene and the number of seams that can be suppressed remains constant.

To implement these cases, rupture detection is applied on the number of vertical seams and horizontal seams to identify the important variation of the seams number. The first step is to apply a median filter on the number of seams that can be suppressed as a function of time to reduce the local variation created by noise or error of salient object detections in some frames. Then, a detection of rupture is applied to segment the sequence into GOPs. The rupture detection is based on the variation of the function compared to its median. A *GOP_Threshold* is used to define when a new segment is created. The number of removable seams is defined by its median value if the GOP's

amount of seams does not change. If the GOP's amount of seams change, the number of removable seams is defined by its minimum value. In this way, the reduction process is improved while preserving the salient objects, especially for the monotonic parts. Finally, the number of suppressible seams in each GOP is rounded to the nearest multiple of 16.

This breakdown detection is performed vertically and horizontally. The combination of these two analyses gives an adaptive cut of the video with different dimensions. Figure 30 resumes this approach. For the horizontal and the vertical seams, the breakdown detection is applied, as illustrated in graphs Figure 30.a and Figure 30.c. Then, for each part, a number of seams multiple of 16 is found as in Figure 30.b and Figure 30.d. By combining the vertical and the horizontal analyses, the GOPs are defined with their dimensions as in Figure 30.e.

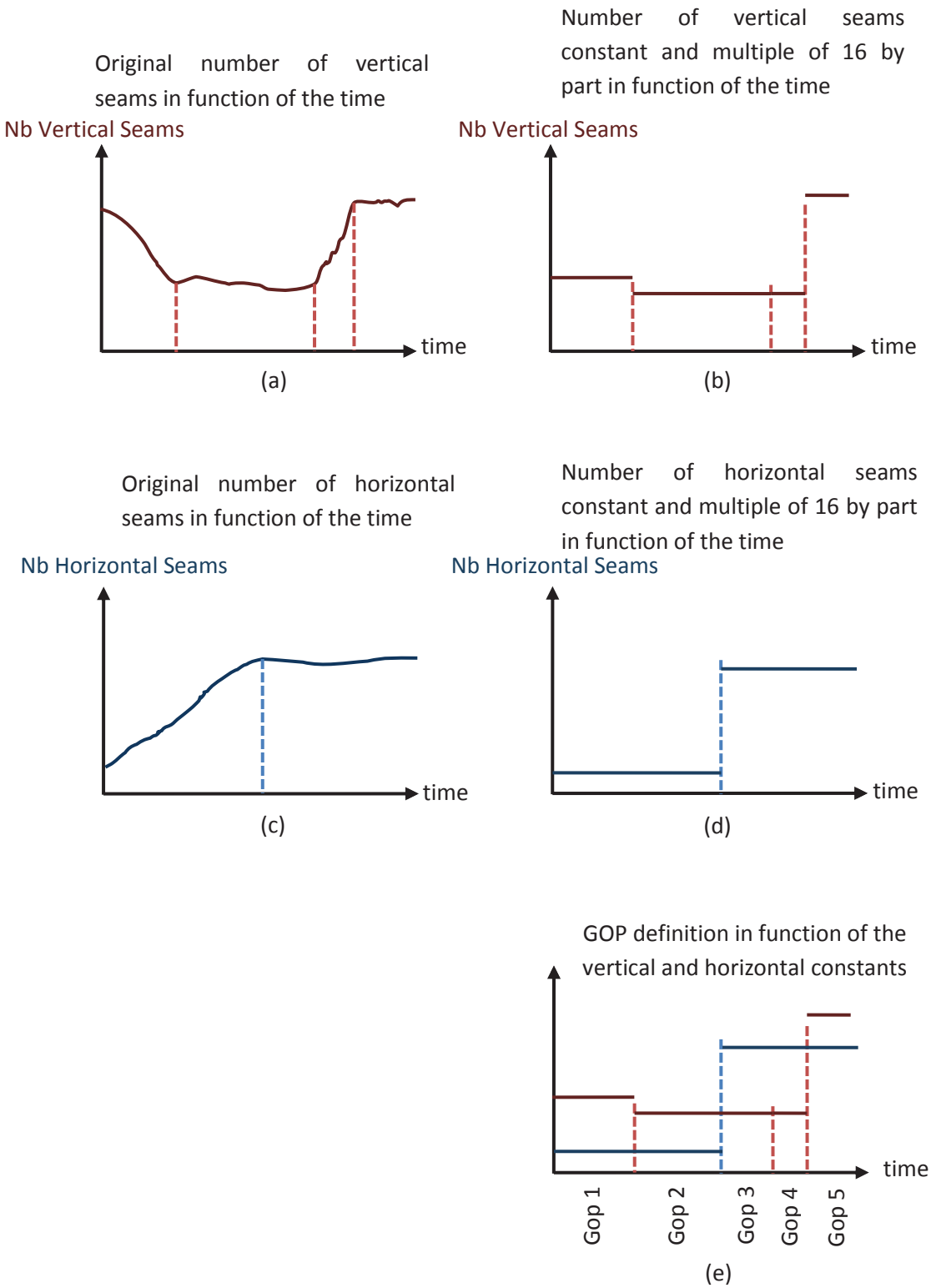


Figure 30 GOP definition depending on the content. Horizontal continue lines represent the number of seams. Vertical dotted lines represent the GOP segmentation. The blue is used for horizontal seams and the red is used for vertical seams

5.4.2. Spatio-temporal seam clustering

After having defined the GOP and the corresponding number of seams that can be suppressed, spatio-temporal clustering is used to identify groups of seams. These groups of seams will be used to model and encode the most important seams and to identify isolated ones. First, as the seam carving is an iterative process, the coordinates of the seams position depend on the reduce image and they can cross one another. So the seams are ordered and expressed in the original images. In this way, the seam position can be used for the clustering and the seams do not cross each other.

To perform the spatio-temporal grouping, the seams are first grouped together with a *Spatial Distance* and a *Spatial Threshold.* The *Spatial Distance SD* is defined as

(12). $\forall j \in [1, J-1],\ SD_{(j,t)}\left(Seam_{(j,t)}, Seam_{(j+1,t)}\right) = \max_{i=1\ldots N}\left(Seam_{(j,t)}(i), Seam_{(j+1,t)}(i)\right)$

where *N* is the length of a seam, i the index inside the seam and J the number of seams suppressed into the frame. For a vertical (respectively horizontal) seam, $Seam_{(j,t)}$ is the horizontal (respectively vertical) position of the *j*-th seam in the frame t and $Seam_{(j+1,t)}$ is the horizontal (respectively vertical) position of the seam *j*+1-th in the frame *t*. This maximum distance has been chosen as it successfully identifies salient objects between seams, contrary to the mean or the median. Then the group of seams are created and numbered.

(13). $Gpe_{Seam(m,t)} = l$

(14). $if\ D\left(Seam_{(j,t)}, Seam_{(j+1,t)}\right) < Sp_{Th} => \ Seam_{(j+1,t)} \in Gpe_{Seam(m,t)}$

(15). $if\ D\left(Seam_{(x,t)}, Seam_{(x+1,t)}\right) \geq Sp_{Th} => \ l = l + 1,\ Seam_{(x+1,t)} \in Gpe_{Seam(m,t)}$

where Sp_{Th} is the *Spatial Threshold* and represents the maximal distance between two consecutives seams found in the same group. It has been experimentally set to 12 pixels. $Gpe_{Seam(m,t)}$ is the *m*-th group of seams in the frame *t*. l is the numerous of the group of seams and is initialized at 1.

Then, the groups are temporally linked together using the symmetric difference between the groups of seams at *t* and *t*+1. The area of the *m*-th groups of seams at the frame *t,* $A\ Gpe_{Seam(m,t)}$ is delimited vertically by the most left seams, *Border seam L*, the most right seams, *Border seam R*, and horizontally by the line *y* = 1 and *y* = N.

(16). $A\ Gpe_{Seam(m,t)} = \int_1^N Border\ seam\ R\ (i) - \int_1^N Border\ seam\ L\ (i)$

The Symmetric Difference *SymDif* between the group $Gpe_{Seam(m,t)}$ and $Gpe_{Seam(k,t+1)}$ is defined as the area of the union of $Gpe_{Seam(m,t)}$ with $Gpe_{Seam(k,t+1)}$minus the area of the intersection of $Gpe_{Seam(m,t)}$ with $Gpe_{Seam(k,t+1)}$

(17). $\forall m \in [1, M], \forall k \in [1, K],$

$SymDif\left(Gpe_{Seam(m,t)}, Gpe_{Seam(k,t+1)}\right) = \left(A\ Gpe_{Seam(m,t)}\ \Delta\ A\ Gpe_{Seam(k,t+1)}\right) =$

$\left(A\ Gpe_{Seam(m,t)}\ \cup A\ Gpe_{Seam(k,t+1)}\right) \backslash \left(A\ Gpe_{Seam(m,t)}\ \cap A\ Gpe_{Seam(k,t+1)}\right)$

Where M is the number of groups of seams for the frame *t* and K is the number of groups of seams for the frame *t*+1.

Then, the temporal regrouping is applied depending on the minimum of the $SymDif$. By this way the numerous of the group of seams at the frame t having the smallest distance with the group of seams at the frame t+1 is given to it. If the $SymDif$ is superior to the *Temporal_Threshold*, a new class is created and if the distance is superior to the *Temporal_Threshold* otherwise, the label of the group of the previous frame is transmitted to the group of the current frame having the smallest distance.

Figure 31 illustrates the spatio-temporal clustering for two consecutive frames. The seams are represented in color. The spatial clustering is first applied, and leads to three group of seams (illustrated with the orange braces). Then the temporal clustering is carried out, and each group of seams of the frame t is compared with the group of seams of the frame t+1. The different arrows illustrate the temporal grouping. The temporal link is done between the groups having the minimum temporal distance.

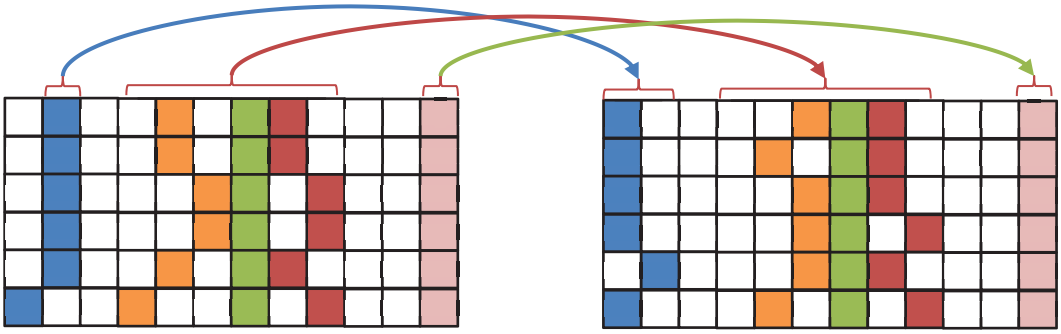


Figure 31 Illustration of the spatio-temporal seam clustering

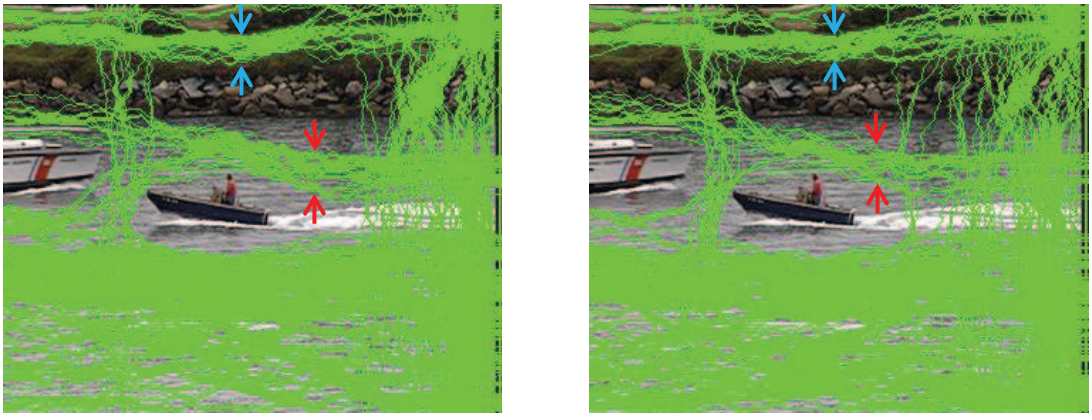


Figure 32 Illustration of seams for two consecutive frames on the coastguard sequence. Arrows show example of group of seams.

On Figure 32, seams are illustrated in green for two consecutive frames on the coastguard sequence. Blue and red arrows show an example of group of seams that are clustered together.

5.4.3. Isolated seam discarding

During the previous step, the number of groups of seams is defined for the GOP. Moreover, for each group, we know the number of seams associated and the number of frames where it is present. As each group will be modeled and encoded, it is important to encode only groups that are important in order to avoid high overhead.

For this purpose, all the small groups, having a percentage of the total number of seams inferior to a threshold *Outliers_Number_Threshold*, are deleted. In addition, groups of seams have to be present in enough frames to be temporally consistent. This is defined with the threshold *Threshold_Outliers_Length*. Therefore, groups that are only present in few frames are also discarded.

On Figure 33 some isolated seams are shown with the red arrows.

Finally, since the number of seams in each frame should be constant throughout the GOP, discarded seams are reallocated among the other remaining groups of seams.

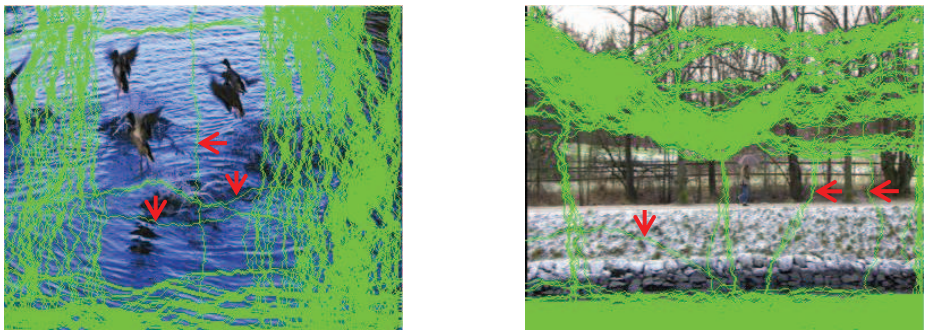


Figure 33 Example of isolated seams on an image of the duck sequence (left image) and on an image of the parkrun sequence (right image). Isolated seams are shown by the red arrows

5.4.4. Group of seams modeling

After having defined the groups of seams, they have to be modeled before being encoded. As we suppose that salient parts of the sequence do not contain seams, the modeling should not modify the outside of the groups of seams while creating the maximum of diversity within the groups of seams.

In this approach, border seams are approximated in a different way that the seams within the group.

Figure 34 illustrates two groups of seams with in red the border seams and in blue the seams included in a group of seams. The first group has 3 seams included between the two borders seams and the second group has one seam included between the two borders seams.

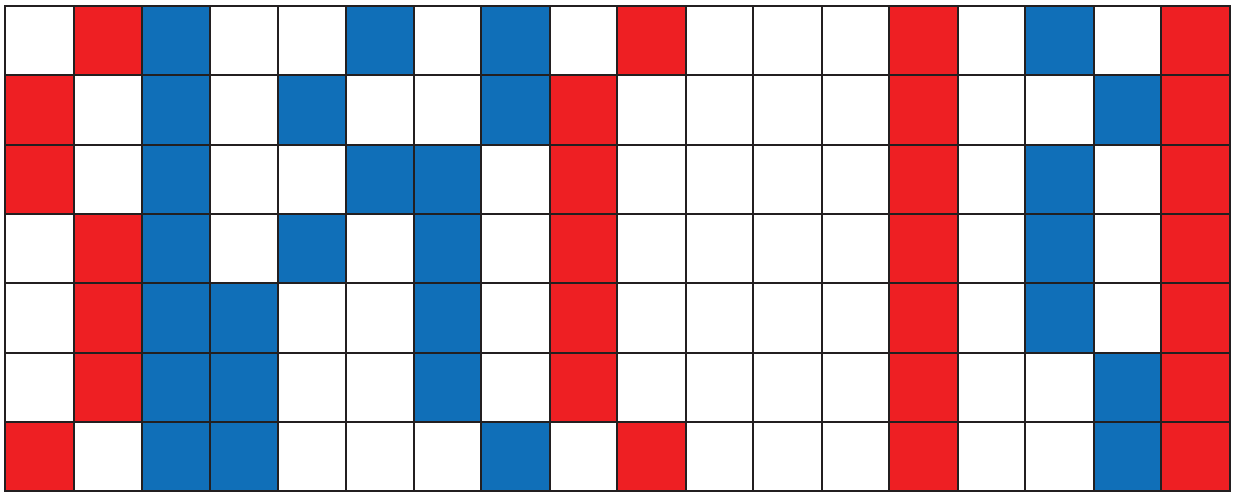


Figure 34 Illustration of two groups of seams with in red the borders seams and in blue the seams included in the group of seams.

Figure 35 resumes the process to encode the borders of groups of seams (red seams in Figure 34). Each border seam will be approximated by a polynomial that needs less information to be encoded. As the outside of the group of seams is considered as salient parts, the polynomial seams have to be totally included inside the group. So all the parts of the left polynomial seam should be at the right of the left border seam and all the parts of the right polynomial seam should be at the left of the right border seam. The first step is the combination between the polynomial seam and the border seam to create a combined seam that is more included into the group of seams. It is done by following the equation.

(18). $\forall i \in [1, N]\ Combined\ seam\ L(i) = \max\big(Polynomial\ seam\ L(i), Border\ seam\ L(i)\big)$

(19). $\forall i \in [1, N]\ Combined\ seam\ R(i) = \min(Polynomial\ seam\ R(i), Border\ seam\ R(i))$

Where N is the length of the seam, *Polynomial seam L* is the left polynomial seam and *Border seam L* the seam border at the left of the group of seam. *Polynomial seam R* is the right polynomial seam and *Border seam L* the seam border at the right of the group of seam. For the first iteration, the polynomial seam is initialized with the border seam.

The combined seam obtained is then approximated by a polynomial seam. Then, we check if the polynomial seam is included into the group of seam :

(20). $if\ \sum_{i=1}^{N}\Big(\big(Polynomial\ seam\ L(i) - Border\ seam\ L(i)\big) > 0\Big) > 0 => not\ included$

(21). $if\ \sum_{i=1}^{N}\Big(\big(Border\ seam\ R(i) - \ Polynomial\ seam\ R(i)\big) > 0\Big) > 0 => not\ included$

If the polynomial seam is not totally included, it is combined another time with the border seam. Otherwise, the shape of the polynomial seams is encoded.

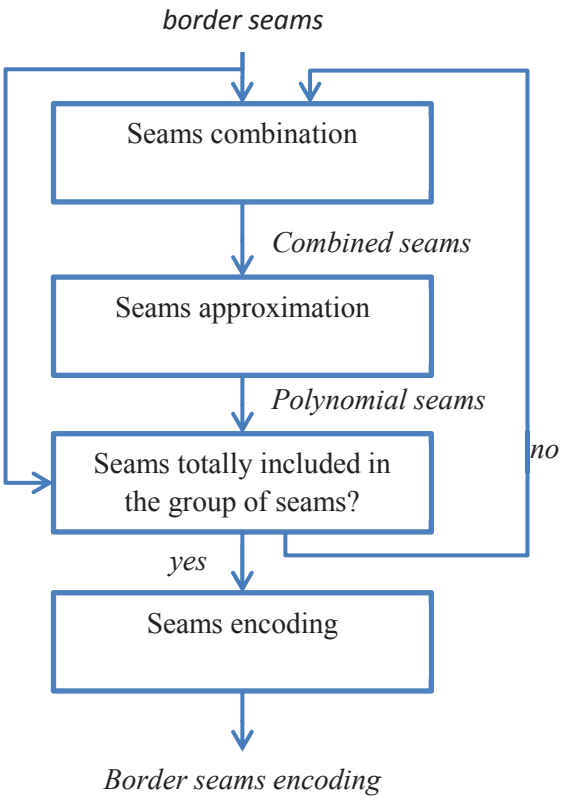


Figure 35 Groups of seams borders approximation and encoding

Figure 36 illustrates the process to obtain the left combined seam in three steps. During the first step, the original border seam is identified and is represented in red. During the second step, the border seam is approximated by a polynomial seam represented in green. The last step is to obtain the combined seam in purple being the most right parts between the original seam and the polynomial seam. To obtain the right combined seam, the process differs only during the combination where the most left parts are selected instead of the most right parts.

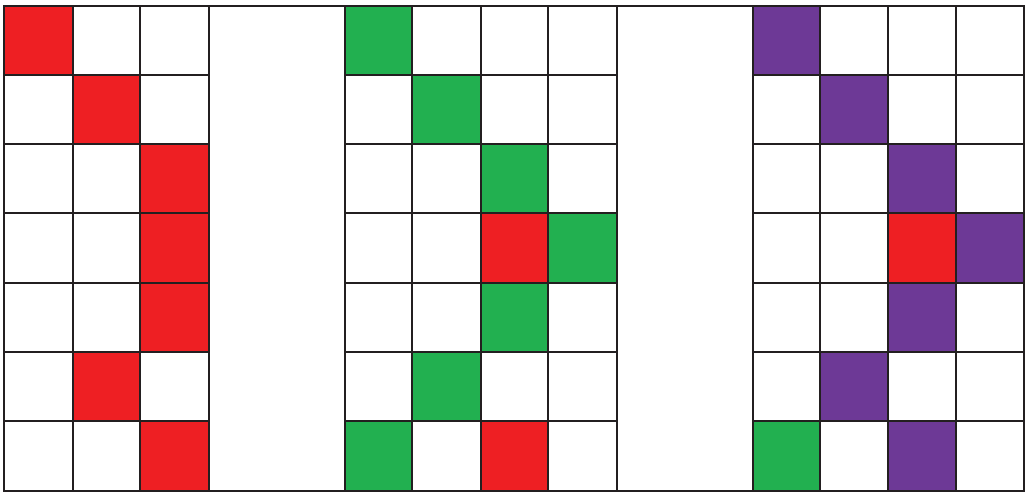


Figure 36 Iterative border seams modeling. First step: original left border seam in red. Second step: original left border seam in red and polynomial seam in green. Third step: original left border seam in red, polynomial seam in green and combined seam in purple.

For the seams included in the group of seams (blue seams in Figure 34), they are modeled by a uniform distribution between the two border polynomial seams.

Figure 37 illustrates the result of the seam clustering, isolated seam discarding and group of seams modeling. It can be seen that 6 vertical groups of seams are identified and an isolated vertical seam is discarded. The shape borders of groups of seams are approximated by polynomial seams and by this way, the quantity of information to transmit is reduced and the outside of the groups of seams is preserved. Horizontally, one group of seams is identified.

Original seams Modeled seams

Figure 37 Illustration of the original seams on the left image and the modeled seams with [Décombas 2014] on the right image for the parkjoy sequence.

5.4.5. Seam encoding

To rebuild the seams at the decoder side, some information has to be transmitted. More specifically, for each frame, the number of seams associated with the groups of seams and data about the seams modeling are sent. The latter includes the border seams approximated by polynomials of degree 3. For a seam of length N, the coefficients of the polynomial are expressed in spatial coordinates by evaluating the position of polynomial seams at x, $x \in [1 \dots N]$. Then a predictive scheme is used.

Figure 38 illustrates with more details the predictive models for the encoding for the group of seams borders. The first coordinate of the first seam in the first frame is fully encoded, and then the next coordinate of this seam is predicted from the previous one. For the next seam, if it is the other border of a group of seams, the first coordinate is predicted from the first coordinate of the previous seam plus the number of seams inside the group of seams, otherwise it is predicted from the previous seams of the previous group. If the next seam is not the other border of a group of seams, the prediction of the first coordinate is just the difference with the first coordinate of the previous seams. After having performed the prediction of the first frame of the GOP, the prediction is done temporally for the others frames.

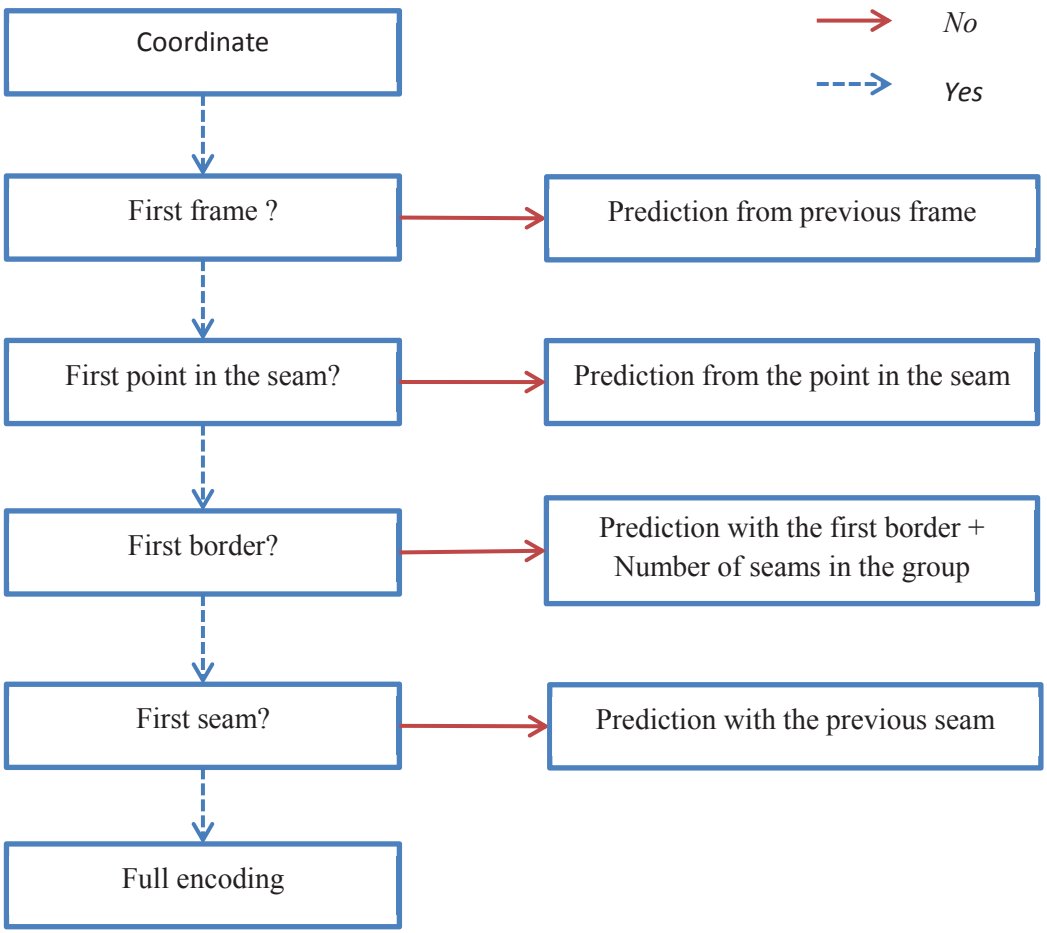


Figure 38 Predictive model for the group of seams borders

Finally, an arithmetic encoding based on the Matlab implementation 2012b (arithenco.m and arithdeco.m) that use [Sayood 2000] is also performed to reduce the quantity of transmitted information.

5.5.Conclusion

In conclusion, schemes to do video compression based on seam carving have been presented. The advantage of using seam carving for video compression is that salient information is concentrated in a reduced resolution sequence and the non-informative background is suppressed, leading to a significant bitrate reduction and a better preservation of the salient parts. Since the seam carving process is not reversible, the correct position of the objects may be lost during the seam reduction. It is therefore necessary to transmit some additional information about the seams in order to properly recover the original dimension of the video sequence and the position of the salient parts.

To represent the seams, two main approaches have been proposed. The first one is based on a encoding of some key points that will be used at the decoder to modify the cumulative energy maps. In these approaches saliency maps are computed at the decoder. These approaches has been published in [Décombas 2011], [Décombas 2012a] and [Décombas 2012b]. The second approach is based on a modeling of the seams at the encoder side and does not need saliency maps at the decoder. It has been published in [Décombas 2014]. In the two approaches, some common innovations have been introduced. The approach in [Décombas 2014] being the most refined one, it will be used to do the evaluation.

6. Evaluation methodology and metrics

6.1. Introduction

The proposed approach to do video compression with seam carving is based on two important steps. The first one is to determine what is important in the image, and the second one is the process of seam carving to reduce the bitrates. Therefore, two evaluations have to be carried out. The first one is to measure the capacity of the saliency maps to determine what is important in the image. The second one is to measure the bitrate saving.

First, the metric to measure the quality of a saliency maps will be presented with eyes tracking reference and with a manual binary mask. We used the term metric by meaning evaluation criterion. Then, we will see the traditional metrics for video compression and for resizing, as well as their limitations. Finally, we will describe a new object based quality metric based on SIFT and SSIM [Décombas 2012c] that introduces two measures. The SSIM_SIFT that measures the compression artifacts around SIFT points and the GEOMETRIC_SIFT that measures the geometric distortion.

6.2. Metrics for saliency maps

The saliency model tries to reproduce the human visual system. The result is a map of probability, with a high probability for the parts that should be interesting and a low probability otherwise. To evaluate the performance of the saliency model, references are needed. Usually, an eye tracking device is used to measure where a candidate is watching and this process is repeated for several candidates. In this way, a reference probability map can be built by using a dilation filter. Another way to evaluate saliency map is by an object detection approach. In this case, a manual segmentation of salient object is needed.

6.2.1. Eye tracking reference

For the human fixations, several models are available, but in our case, we have concentrated our evaluation based on three different metrics

The first one is the Area Under the ROC curve (AUROC) [Lau] which focuses on saliency location (on the saliency map) at gaze positions (on the heatmap). The objective is to obtain a high score that indicates good performance. The Normalized Scanpath Saliency (NSS) [Borji 2011] focuses on saliency values (on the saliency map) at gaze positions (on the heatmap). As the AUROC, a high score mean good results. KL-Divergence [Mancas] focuses on the discrepancy between saliency and gaze probability distributions. Conversely of the two previous metrics, low scores are better for KL-Divergence.

To compute the Area under the ROC curve [Lau], fixations pixels are counted once and the same number of random pixels are extracted from the saliency map. For one given threshold, saliency pixels can be treated as a classifier, with all points above threshold indicated as 'fixation' and all points below threshold as 'background'. For any particular value of the threshold, there is some fraction of the actual fixation points which are labeled as True Positives (TP), and some fraction of points which were not fixation but labeled as False Positive (FP). This operation is repeated one hundred times. Then the ROC curve can be drawn and the Area Under the Curve (AUC) computed. An ideal score is one while random classification provides a value of 0.5.

The Normalized Scanpath Saliency (NSS) metric was introduced in 2005 by Peeters and Itti and an implementation is available in [Borji 2011]. The idea is to quantify the saliency map values at the eye fixation locations and to normalize it with the saliency map variance:$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}}$ where p is the location of one fixation and SM is the saliency map which is normalized to have a zero mean and unit standard deviation. Indeed, the NSS score should be decreased if the saliency map variance is important or if all values are globally similar (small difference between fixation values and mean) because it shows that the saliency model will not be very predictive, while he will precisely point a direction of interest if the variance is small or the difference between fixation values and mean high.

The NSS score is the average of NSS(p) for all fixations:

(22). $NSS = \frac{1}{N} * \sum_{p=1}^{N} NSS(p)$

The Kullback-Leibler divergence is a commonly used metric to estimate an overall dissimilarity between two distributions. Many authors like [Tatler 2005], [Rajashekar 2004], [Le Meur 2007] already used this metric to compare saliency maps with human eye fixations. The KL-Div is a measure of the information lost when the saliency maps probability distribution (SM) is used to approximate the human eye fixation map probability distribution (FM).

(23). $KL_{div} = \sum_{x=1}^{X} FM(x) * \log(\frac{FM_{Norm}(x)}{SM_{Norm}(x) + \varepsilon} + \varepsilon)$

where X is the number of pixels and ε is a small constant to avoid log and division by zero. SM and FM distributions are both normalized as in the equations below

(24). $SM_{Norm}(x) = \frac{SM(x)}{\sum_{x=1}^{X} SM(x) + \varepsilon}$,

(25). $FM_{Norm}(x) = \frac{FM(x)}{\sum_{x=1}^{X} FM(x) + \varepsilon}$.

When the two maps are strictly equal, the KL-divergence value is zero.

In this way, the performance of the saliency models can be evaluated.

6.2.2. Manual binary mask

To do the evaluation of the capacity of your approach to well detect the objects, manual binary mask has been realized in YUV sequences. These masks will be used as reference and are presented in Sec.7.2. Precision – recall metric [Rijsbergen 1979] is used on the binary mask and a binarization of the saliency map to measure the performance. This metric is based on true positives (*tp*), true negatives (*tn*), false positives (*fp*) and false negatives (*fn*) that compare the predicted results with the reference results. The Table 5 illustrates their definition

	Reference results	
Predicted results	*tp* : Correct result	*fp* : unexpected result
	fn : Missing result	*tn* : Correct absence of result

Table 5 Precision / recall table.

The precision is the number of relevant points compared with the total number of points found and is defined as:

(26). $Precision = \frac{tp}{tp+fp}$.

The recall is the number of relevant points compared with the total number of important points in the reference. It is defined as*:*

(27). $Recall = \frac{tp}{tp+fn}$.

To link the Precision and the Recall, the F-measure is usually used and it is defined as:

(28). $F = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall}$

6.3.Metric for compression and resized images

6.3.1. Introduction

After having defined and evaluated what is important in the image, the second step of our approach is to suppress the non-important parts of the sequence to reduce the bitrate. An evaluation of the quality compared with the bitrate is usually done in compression. First, traditional metric will be presented, then some work trying to evaluate images resized or having small distortion will be detailed. Then the proposed metric in [Décombas 2012c] will be exposed.

6.3.2. Traditional image quality metrics

Traditional fidelity metrics such as Peak-Signal-to-Noise-Ratio (PSNR) or Structural SIMilarity (SSIM) [Wang.Z 2004] compare corresponding pixels or blocks in the reference and processed images.

The PSNR is defined as:

(29). $PSNR = 10\ .\log_{10}(\frac{d^2}{MSE})$

with d the dynamic of the image and the Mean Square Error (MSE) defined as:

(30). $MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\|I_{ref}(i,j) - I_{eval}(i,j)\right\|^2$

with m, n the dimension of the image, I_{ref} the reference image and I_{eval} the image to evaluate.

The SSIM is computed on windows of size N x N. For a windows (x,y), the SSIM is defined as:

(31). $SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2cov_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)\ (\sigma_x^2+\sigma_y^2+C_2)}$

with μ_x the mean of x, μ_y the mean of y, σ_x the variance of x, σ_y the variance of y, cov_{xy} the covariance of x, y, $C_1 = (k_1 d)^2$, $C_2 = (k_2 d)^2$two variables to stabilize the denominator when it is very small. d is the dynamic of the image, $k_1 = 0.01$ and $k_2 = 0.03$.

These two traditional metrics evaluate all the pixels of the images. They are sensitive to geometric deformations, background suppression or synthesis and do not manage bitrate allocation in function of the regions. Another problem is that these two metrics cannot be used to evaluate two images with different sizes.

6.3.3. Metrics for images with different resolutions

We remind that our goal is to develop a full reference object-based visual quality metric to evaluate a semantic coding system under the assumption that the position and the shape of objects may have been considerably modified.

In [Wang.Z 2005], Wang and Simoncelli propose a complex wavelet domain image similarity measure that is insensitive to luminance change, contrast change and spatial translation. This metric is robust for small geometric distortions relative to the size of the wavelet filter. However, it does not handle large displacements, nor assesses geometric deformations.

Rubinstein *et al.* presents a subjective evaluation for image retargeting and intents to create an objective metric [Rubinstein 2010]. Specific features are identified in retargeted media that are more important for viewers. It is concluded that the resizing method having the best subjective score is also the one having the worst score with their objective metric. Therefore, a reliable metric remains a challenge.

A full reference metric based on Scale-Invariant Feature Transform (SIFT) [Lowe 2004] and SSIM has been developed by Azuma *et al.* [Azuma 2011] in order to evaluate images resized by different retargeting algorithms. In [Liu 2004], Liu *et al.* present an objective metric simulating the Human Vision System (HVS) based on global geometric structures and local pixel correspondence based on SIFT.

The common objective of [Azuma 2011], [Liu 2004] is to evaluate resized (smaller) images. However, both methods are not designed to measure compression artifacts and do not take into account geometric deformations possibly occurring in content-based coding schemes (e.g. [Décombas 2011], [Tanaka 2010b]). Another limitation is that both metrics compare two entire images. Therefore, they fail to assess the quality of a specific (salient) object. Finally, the metric in [Azuma 2011] has not been validated by subjective tests.

6.4.Proposed object based quality metric based on SIFT and SSIM

6.4.1. Introduction

In this section, we introduce an object-based visual quality metric by selecting and matching SIFT points in the object to evaluate. This metric is not a mathematic metric because it needs to satisfy four properties: non-negativity, identity, symmetry and triangular inequality. The triangular inequality is easy to prove due to the complex dependence on SIFT. So we used the term metric by meaning evaluation criterion. SIFT points allow to put in correspondence the same object in the reference and processed images even though it has been geometrically distorted. Our proposed metric gives two scores. The first one applies SSIM in the neighborhood of matching SIFT points and

is referred to as SSIM_SIFT. Thus, it measures traditional compression artifacts such as those resulting from H.264/AVC (Advanced Video Coding). The second score measures the geometric deformation of the object. It is based on the standard deviation of matching SIFT points coordinates and is referred to as GEOMETRIC_SIFT. These two measures have been validated by subjective evaluation following the Double Stimulus Impairment Scale (DSIS) protocol [ITU-R 2009].

In summary, our main contribution is a metric called SSIM_SIFT and detailed in [Décombas 2012c] that (7a) is based on a combining of SSIM and SIFT, has the advantage (7b) to be insensible at the background suppression and synthesis and allows to (7c) measure compression artifacts due to traditional encoders like H.264/AVC (SSIM_SIFT) and the geometric deformation of the objects (Geometric_SIFT). Subjective validation has been done showing its efficiency.

6.4.2. Problems definition

For the purpose of object-based semantic coding, approaches based on seam carving have been proposed by Tanaka in different articles like in [Tanaka 2010b] and us like in [Décombas 2011]. Seam carving is a method of resizing that suppresses or adds lines in non-salient parts of an image. Hereafter, we more specifically consider the previously presented method [Décombas 2011] without loss of generality. In this method, seam carving is applied as pre- and post-processing in conjunction with a conventional H.264/AVC video coding scheme, as illustrated in Figure 39 . Consequently, the method may introduce both traditional compression artifacts of H.264/AVC and geometric artifacts from the seam carving and synthesis.

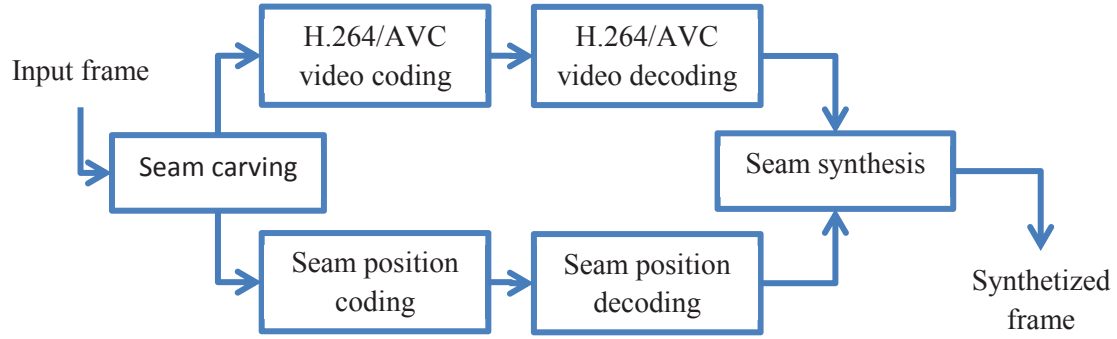


Figure 39 Architecture of seam carving based video coding

In this context, where salient objects are preserved but may undergo small displacements and deformations and where the background can be strongly modified and synthesized, traditional quality metrics such as PSNR or SSIM [Wang.Z 2004] fail.

6.4.3. SSIM_SIFT and GEOMETRIC_SIFT metrics description

We propose a full reference metric to assess an object based coding system which has possibly modified the position and/or the shape of objects.

The metric relies on the combination of SIFT and SSIM to evaluate both compression artifacts and object deformations. SIFT [Lowe 2004] is an approach for detecting and extracting local feature descriptors invariant to different changes, in particular rotation, scaling and, in general, geometric deformations. In the proposed metric, SIFT allows to match an object from the original image with a potentially deformed object in the processed image.

Figure 40 represents the proposed full-reference metric. It takes three images as input: the original image, the processed image #1 which has been altered by seam carving but without compression, and the processed image #2 which has been modified both by seam carving and compression.

In a first step, we extract SIFT points from the original image, as well as the processed image #1. This is done in order to avoid the sensitivity of SIFT to coding artifacts. For the reference image, we only select SIFT points inside the considered object.

Next, SIFT points matching is performed from the original image towards the processed image as well as from the processed image towards the original one to increase robustness. Finally, a statistical analysis is performed on the matching distances in order to eliminate outliers. This step is useful to identify erroneously matched SIFT pairs. Formally, a pair of points is considered an outlier if the following equation holds $|D(i) - \mu| > 3\sigma$ with

(32). $\mu = \frac{1}{N}\sum_{i=1}^{N} D(i)$ and

(33). $\sigma = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(D(i) - \mu)^2\right)}$

where *N* is the number of SIFT points, *i* =1,.., *N* denotes the index of the current SIFT point, *D* is the distance between a pair of matching SIFT points in the original and processed images respectively, μ is the mean distance between the matching SIFT pairs, and σ is the standard deviation.

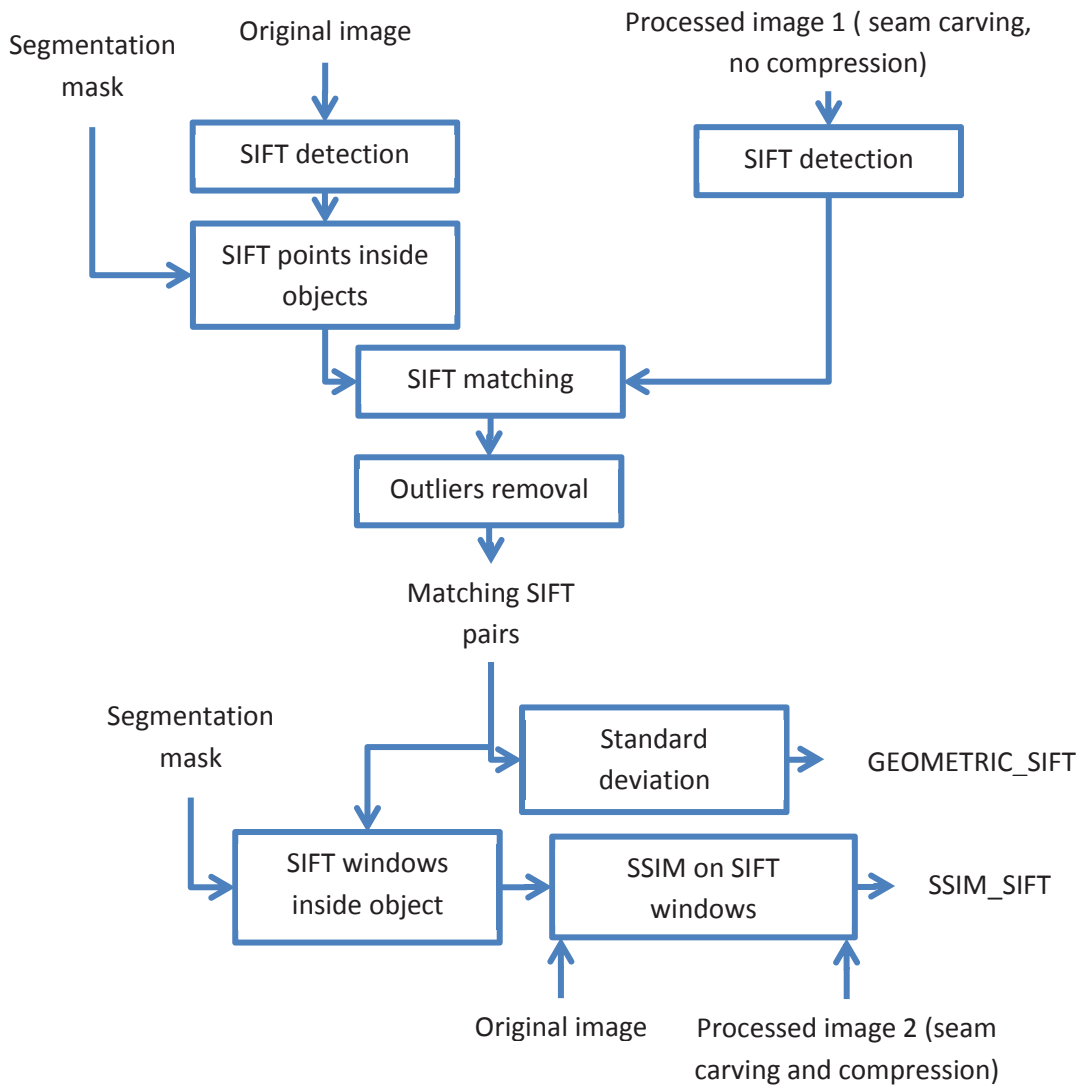


Figure 40 Object-based quality metric. Comparison of the Original image with the Processed image 1 to obtain GEOMETRIC_SIFT score and of Original image with the Processed image 2 to obtain SSIM_SIFT

For GEOMETRIC_SIFT, we simply measure the standard deviation σ between matched SIFT points. This component of the metric captures the non-rigid deformation of the object.

In turn, the SSIM_SIFT component of the metric assesses the visual content of the object. First, non-overlapping *W* x *W* pixel windows are defined centered at each SIFT point and wholly contained inside the object. For this purpose, SIFT points with an associated window laying partly outside the object or with a spatial distance inferior to *W* pixels are discarded. The window dimension *W*=11 is chosen to cover enough of the surroundings but can be defined with a smaller value if the object is small.

Since SIFT points coordinates are not integer values, it can cause a mismatch of ±1 pixel, horizontally and/or vertically, when the window from the original image is compared with the window in the processed image. Thus, nine positions, representing all {-1, 0, 1} pixel shifts horizontally and

vertically, are tested and the one with the minimal Mean Square Error (MSE) is kept. Finally, SSIM is applied on all the windows defined by the above process, leading to the SSIM_SIFT measure.

6.4.4. Subjective test

Following the ITU-R BT.500-12 recommendation [ITU-R 2009], the protocol DSIS is chosen for subjective evaluation.

During the session, as a variation to standard DSIS, the assessor is first presented with a binary mask defining the object, then an unimpaired reference, and finally with the same picture impaired. At the beginning of each session, a training is given to the observers about the subjective assessment. In particular, assessors are specifically instructed to concentrate on the corresponding object. Afterwards, the assessor is asked to vote using the five-grade impairment scale: 5 imperceptible, 4 perceptible, but not annoying, 3 slightly annoying, 2 annoying, 1 very annoying.

Each assessor evaluates 30 images altered with different levels of artifacts spanning a large range of visual quality. The five first images are used for training and corresponding scores are discarded. Subjective scores are then processed and analyzed according to [ITU-R 2009].

6.4.5. Results

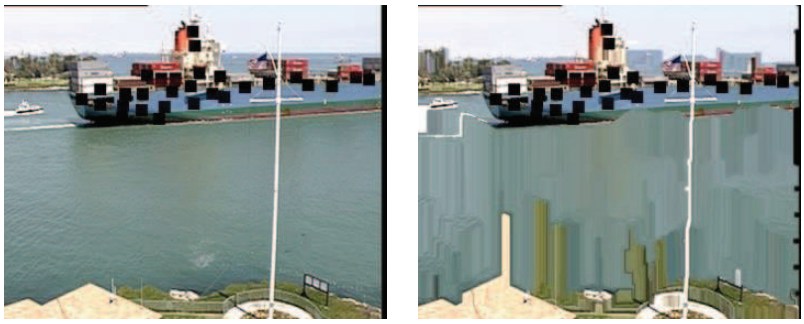
To validate our metric, we use the object-based compression method described in [Décombas 2011] to generate sequences presenting H.264/AVC compression artifacts, geometrical deformations and repositioning artifacts of the salient objects. Experiments are carried out using the test sequences Container and Coastguard in CIF format.

6.4.5.1. Performance of SSIM_SIFT

In a first set of experiments, we evaluate the performance of the proposed SSIM_SIFT, and in particular its ability to assess object-based visual quality in the presence of small displacements and deformations of the salient object.

Seam carving usually stops when it reaches objects defined by a saliency map. In this first experiment, seam carving parameters in [Décombas 2011] have been selected in order to achieve minor geometric distortions of the salient object. Nevertheless, a few seams may go through a salient object leading to artifacts. In addition, small deformations may also be introduced when reinserting seams during synthesis.

Figure 41 illustrates the proposed SSIM_SIFT metric. The Figure 41 (a) and (b) show the SIFT windows (black squares) used to compute SSIM_SIFT in the original image and the processed image #1 (i.e. altered by seam carving but without H.264/AVC compression). In Figure 41 (b), it can be observed that the container ship is well preserved, although the background is noticeably distorted. Moreover, the position and shape of the ship have been slightly altered.



(a) (b)

Figure 41 SIFT windows used to compute SSIM_SIFT (black squares), (a) original image, (b) processed image #1.

As a reference, SSIM_Mask is a straightforward extension of SSIM computed on a salient object as defined by its binary mask. More precisely, SSIM is only calculated on the 11 x 11 windows which are wholly contained in the binary mask.

To validate the proposed SSIM_SIFT metric, a subjective evaluation was done with 14 non-expert assessors following the procedure described in Sec. 4. Six images quantized with QP={18, 36, 39, 42, 48}, for a total of 30 images, were shown in a random order to each assessor. We have found no outliers among assessors when following the procedure defined in ITU-R BT.500-12 [ITU-R 2009].

Figure 42 shows the proposed SIFT_SSIM as a function of the Mean Opinion Score (MOS). The Spearman correlation is 0.86 and the Pearson correlation is 0.86 for the proposed SIFT_SSIM, showing a strong correlation. In comparison, the Spearman correlation is 0.20 and the Pearson correlation is 0.14 for SSIM_Mask. Clearly, SSIM_Mask fails as it cannot handle small geometric displacement or deformation.

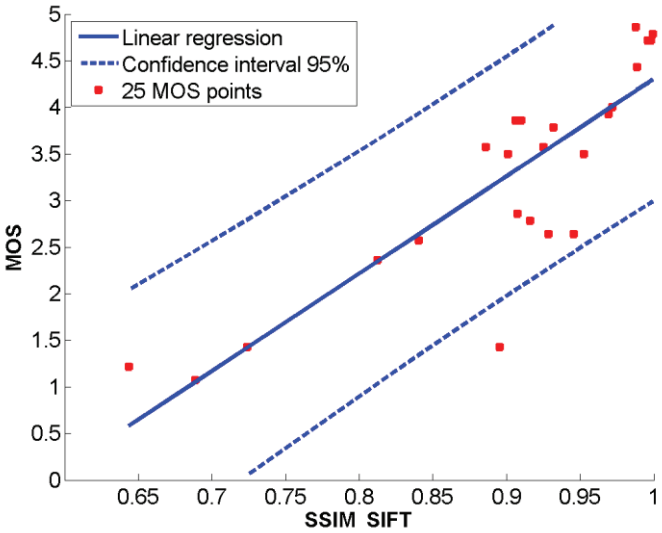


Figure 42 SIFT_SSIM as a function of MOS.

6.4.5.2. Performance of GEOMETRIC_SIFT

We now evaluate the performance of the proposed GEOMETRIC_SIFT to measure object deformation. For this purpose, images with different levels of geometric deformation resulting from seam carving, but without compression artifacts (QP=0), are considered. A new evaluation with 11 assessors has been performed. During this evaluation, no assessor has been detected as an outlier.

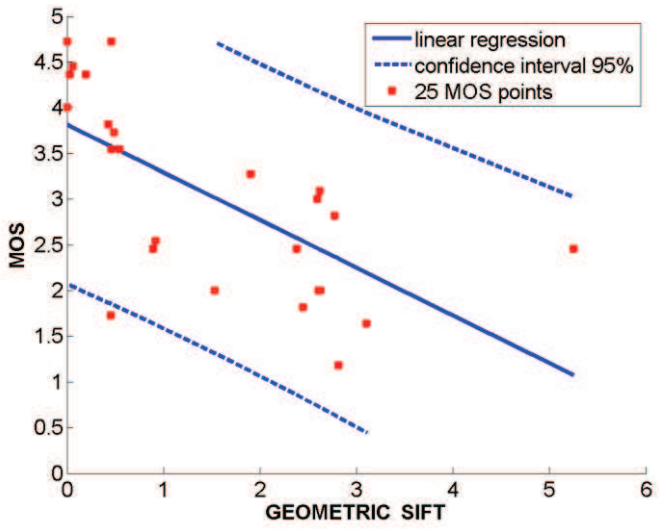


Figure 43 GEOMETRIC_SIFT as a function of MOS.

The result of the experiment is given in Figure 43. The Spearman correlation is -0.74 and the Pearson correlation is -0.67.

Correlations are lowered due two images with poor performances. The image corresponding to GEOMETRIC_SIFT=5.24 (right most point in Figure 43) is shown in Figure 44(a). GEOMETRIC_SIFT is high, as the ship is elongated and has slightly moved as shown in Figure 44 (b). However, as the artifact of translation and deformation is hard to notice, the MOS remains high.

(a) (b)

Figure 44 Container, frame 18, (a) container object, (b) difference between the original and processed images.

Figure 45 Container, frame 23, container object.

The image corresponding to GEOMETRIC_SIFT=0.46 and MOS=1.72 (lower left in Figure 43) is shown in Figure 45. The object of interest (the container ship) itself is well-preserved, however the borders of the object have been strongly distorted. In such a case, the assessor may have evaluated the border region instead of the ship alone. This underlines one of the limitations of this evaluation. The assessor can be influenced by the background.

6.5.Conclusion

In this chapter, the evaluation methodology and the metrics have been presented. As our works combined problematic from computer vision and video compression, it has been necessary to evaluate them separately. To evaluate the saliency maps quality, three metrics has been used. The Area under the ROC curve [Lau] focuses on saliency location at gaze positions, the Normalized Scanpath Saliency (NSS) [Borji 2011] focuses on saliency values at gaze positions. KL-Divergence [Mancas] focuses on the discrepancy of saliency and gaze distributions. For AUROC and NSS, high scores indicate better performance. Conversely, low scores are better for KL-Divergence. On these three metrics, the reference has been realized with an eye tracking device. As one of the proposed saliency models is object oriented, the Precision – recall metric [Rijsbergen 1979] has been used with manual binary segmentation as references. For the video compression evaluation, the traditional metrics does not work because some parts of the video are suppressed. An object-based full reference visual quality metric based on SIFT and SSIM has been proposed in [Décombas 2012c]. It can be used for images where the objects have their position and/or shape modified. The two proposed components have been validated by a subjective evaluation following DSIS. The database for the evaluation is composed of images from coastguard and container with different levels of compression artifacts and geometric deformations. SSIM_SIFT gives a Spearman and a Pearson correlation of 0.86. Evaluation of deformation artifacts with GEOMETRIC_SIFT gives a Spearman correlation of -0.74 and a Pearson Correlation of -0.67. The interest of this metric is it can be used to measure the quality of an object when parts of the background have been suppressed or when the position of the objects has changed.

7. Performance evaluation

7.1.Introduction

As our approach exploit different fields of research, different performance assessment methodologies and metrics are necessary. In function of the field, databases to do the evaluation change. We will first introduce the new database which is used to measure video compression artifacts, object detections and eyes tracking modeling. Then the results for saliency will be presented. We will finish by analyzing the result of video compression by seam carving.

7.2.The database

To evaluate coders, the input sequences have to be in a raw format. The traditional format is the .yuv, and the available sequence can be found in [Xiph.org]. As the metric [Décombas 2012c] needs a manual binary mask, only a few sequences are publicly available with a segmentation mask. Therefore, we have realized several binary masks to increase the database, as reported in [Riche 2014].

To evaluate the saliency model the previous database has been completed with the work of Hadizadeh *et al.* [Hadizadeh 2012]. The eyes tracking results has been realized with 15 independent viewers on a set of 12 standard CIF video sequences. The heat map represents the results of the eye tracking. In Figure 46, example of the original sequence, the manual binary mask and the heatmap are shown.

BUS

CITY

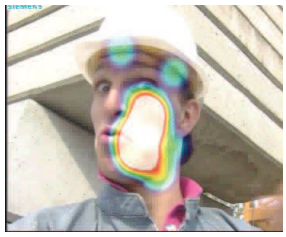
CREW

FLOWER



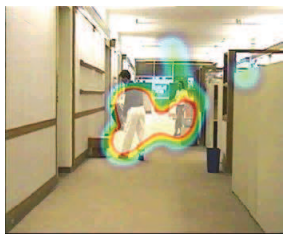
FOREMAN





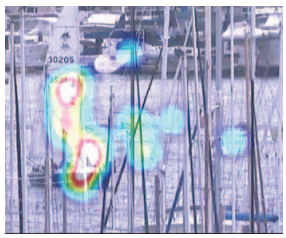
HallMonitor





HARBOUR





MOBILE



mother-daughter



SOCCER



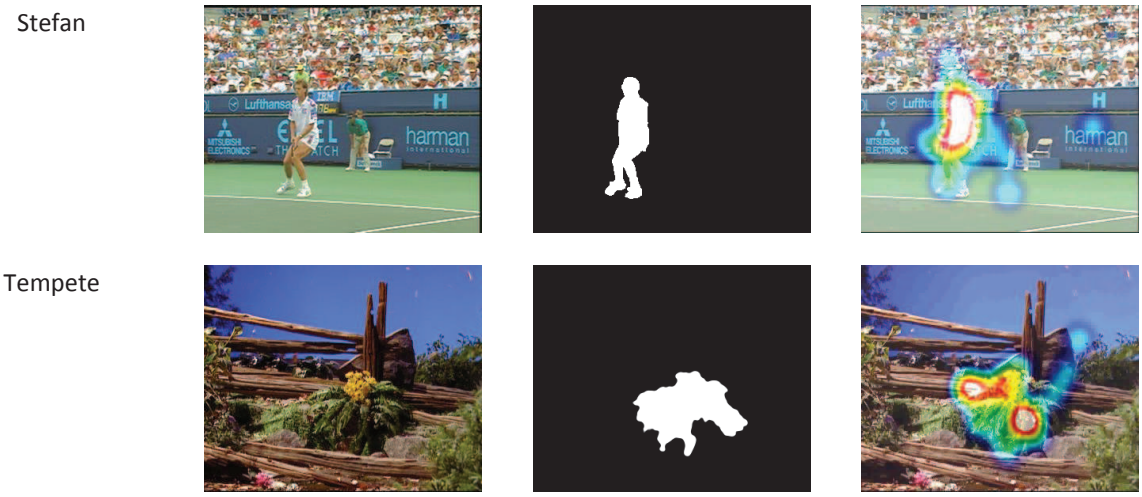


Figure 46 Binary mask available for YUV sequences. From the left to the right: Original image, manual binary mask proposed in [Riche 2014], heatmap from [Hadizadeh 2012].

The experience of eyes tracking has been done twice, to have a first result without knowledge of the sequences, and a second one with viewers knowing what is happening in the sequence. Figure 47 illustrates the available eyes tracking results for the first visualization and the second one. On the fourth image, the distance between the first and second fixation is shown. It can be seen that there is an important difference between the two visualizations.

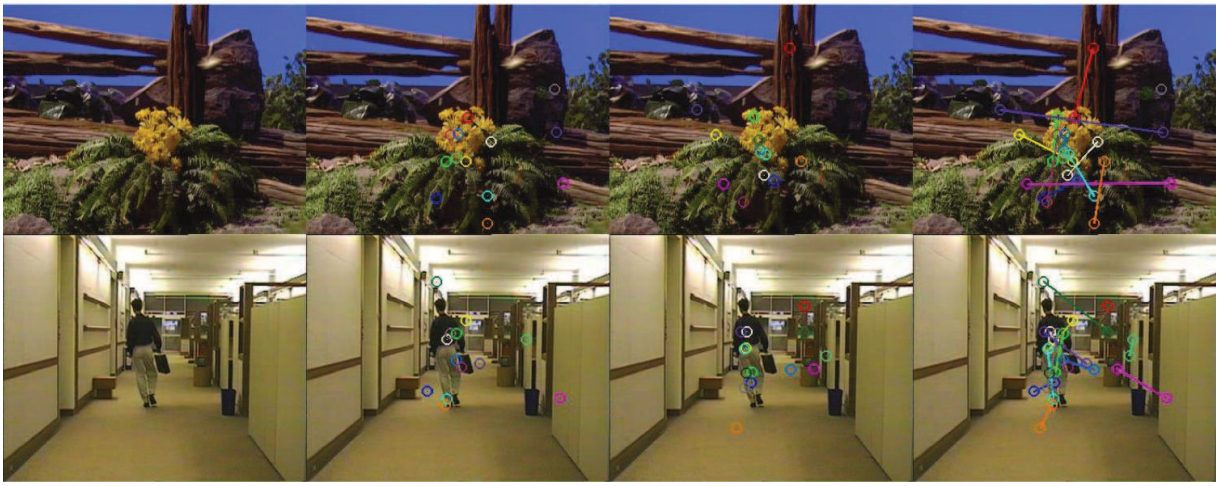


Figure 47 Visual results of the eyes fixations. From the left to the right: Original image, first fixation, second fixation, distance between first and second fixation.

7.3. Results of the saliency approach

To validate the results of the STRAP model [Riche 2014] (see Sec. 4.3), the database presented in Sec.7.2 is used. First a qualitative visualization will be done and then different metrics will be used for validation.

7.3.1. Qualitative validation

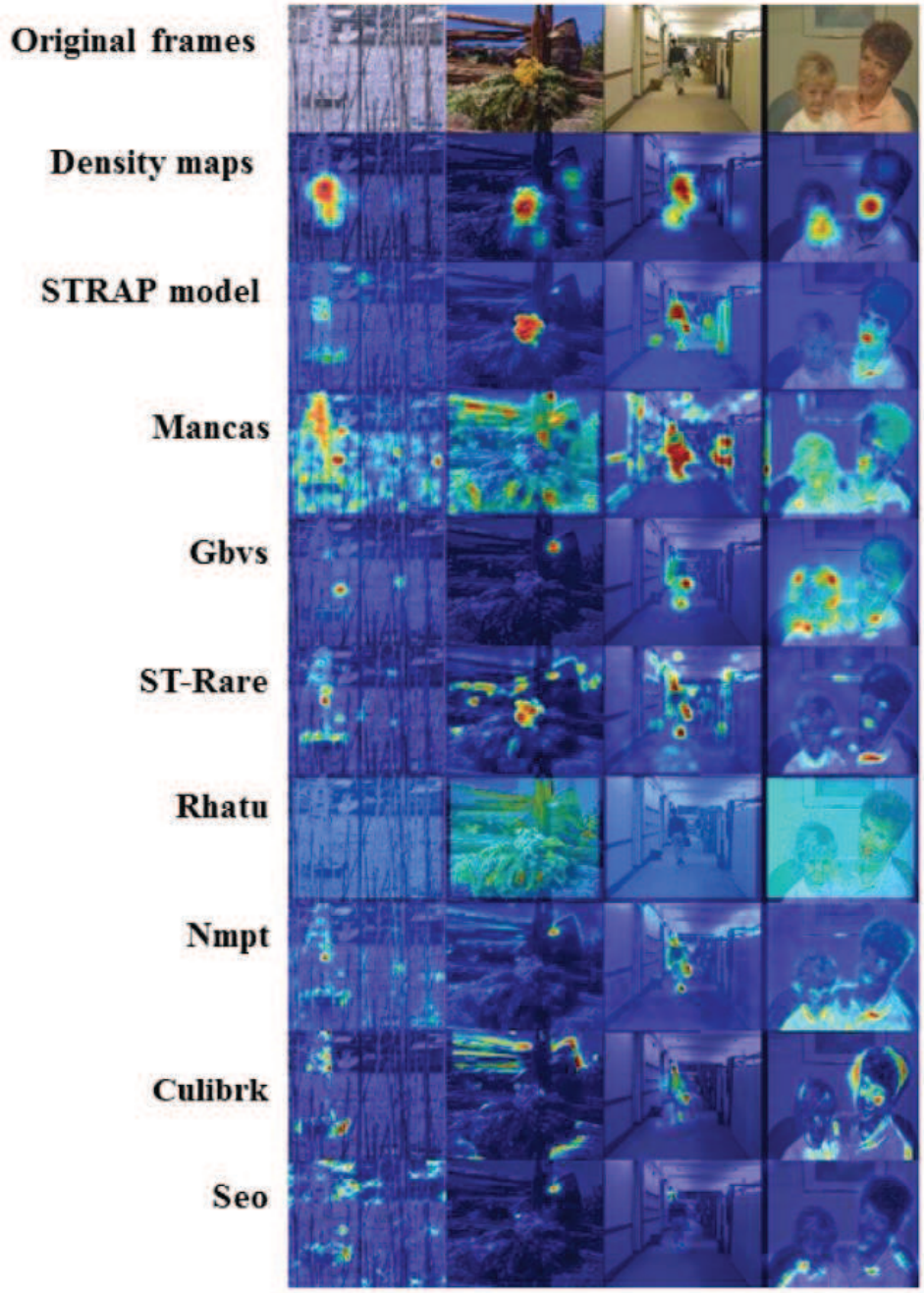


Figure 48 Visual results of the saliency models.

On Figure 48 4 different sequences (Harbour, Tempete, Hall Monitor and Mother Daughter) have been selected to illustrate the results (columns). The two first sequences have lots of movements while the two others sequences have a static camera with a salient moving object. The original frames can be seen on the first line and the density maps, used as reference, are displayed on the second line. The results for the 8 saliency models are illustrated on the others lines. It can be seen that the result for STRAP works very well for Tempete and Hall Monitor. On Harbour and Mother Daughter, the salient objects are well identified but the results could be more intense. For Mancas, which has no camera motion compensation, it is very sensible for Harbour and Tempete where the camera is moving but works well for Hall monitor which has the salient object moving. For Mother

Daughter, it is not very selective because it uses only temporal features, but the faces are overall well detected.

GBVS is a very selective model which provides relatively good highlight of a small part of the salient objects. ST-Rare works also relatively well on the first three sequences, while in Mother and Daughter, the results are bad. In Rathu, the background mean saliency remains relatively important due to the lack of color contrast, leading to results which are not selective enough. NMPT does not work at all in case of complex movements like in Tempete and is weak in moving camera for Harbour. The two other sequences provide average quality results. Culibrk works relatively well in case of a still camera, but the results are disappointing on the sequences where the camera is moving. Seo seems to be the less convincing on these four video sequences with bad results even on the Hall sequence which is probably the easiest.

7.3.2. Quantitative validation

For quantitative assessment, four different experiments are carried out. First, the usefulness of high-level priors is investigated. Then, the proposed STRAP [Riche 2014] model is compared with state-of-the-art techniques for the first viewing salient object detection by using the manual binary segmentation masks as references. Next, a similar comparison is presented for first viewing human prediction, using eyes fixations as references. Finally, a comparison of first viewing and second viewing human predictions is provided.

7.3.2.1. Validation 1: a study of the usefulness of high levels priors

In this first experiment, only two video sequences are used due to the fact that only Foreman and Mother contain detectable faces. It is also important to note that the end of the Foreman sequence does not contain faces. Figure 49 shows the influence of high-level priors (no priors, each prior used separately and the two combined). It can be seen that the model is more efficient when adding high-level priors. In the Foreman sequence, the area at the bottom right of the image is highlighted by spatial features when no priors are considered. The same area is suppressed by the importance of the face and/or the center when high-level priors are added. Moreover, specific salient parts of the face, such as the eyes and the mouth, have also been highlighted. In the Mother sequence, with high-level priors, the top of the collar has lost its importance, whereas the face has become more significant, especially the mouth.

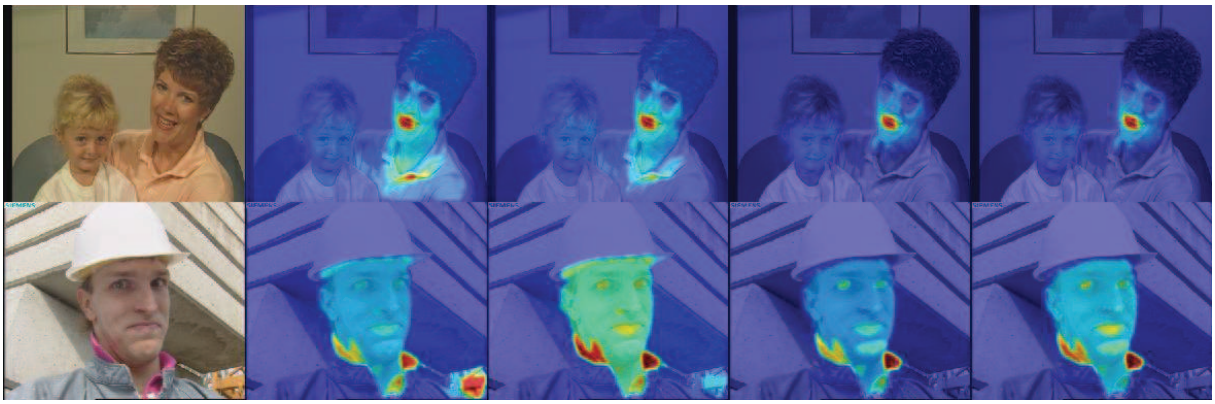


Figure 49 Visual results for different high levels priors. From the left to the right: Original Image, No high priors model, centered Gaussian model, Face detection model, centered Gaussian and face detection models.

In Figure 49, the influence of high-level priors is shown (no priors, each prior used separately and the two combined). A heatmap is used to illustrate the results. The most salient parts are in red and the less one are in blue. It can be seen that the model is more efficient when adding high levels priors. We can see on the Foreman sequence that the static part in right bottom of the image has been suppressed from the saliency map due to the importance of the face and/or the center. It can also be observed that specific saliency parts of the face like the eyes and the mouth have been highlighted. On the Mother sequence, it can be noticed that the top of the collar has lost its importance and the face is more important especially the mouth.

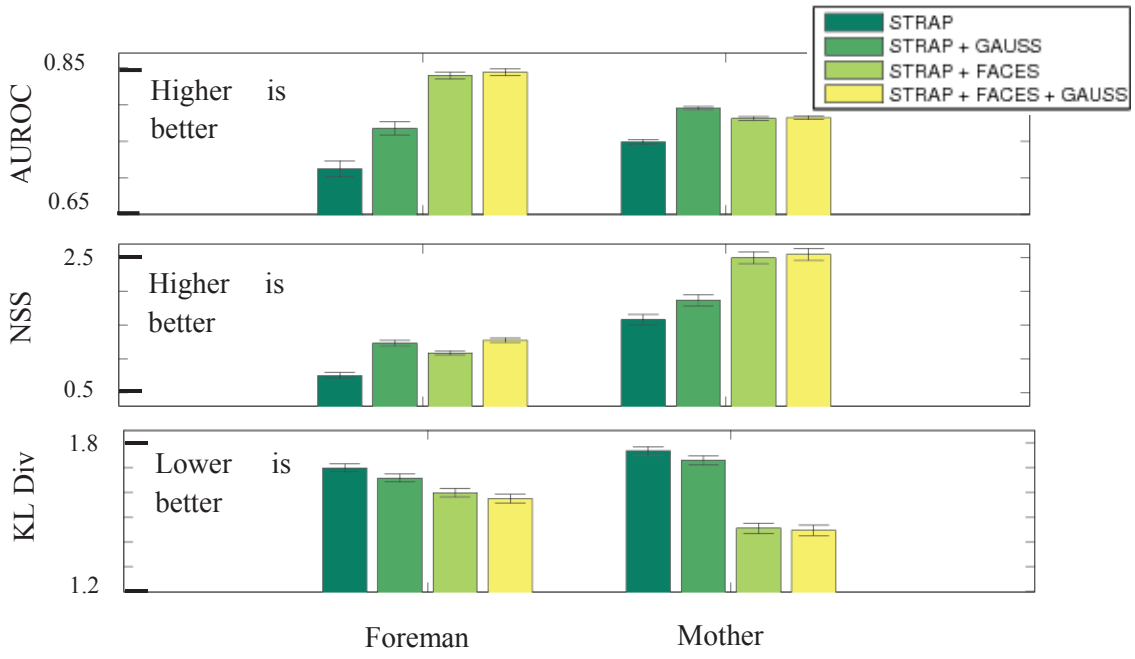


Figure 50 Validation of the usefulness of high level priors. Comparison of STRAP without priors, STRAP + centered Gaussian, STRAP + faces detection and STRAP + faces detection + centered Gaussian.

Figure 50 assesses the performance of high-level priors in the proposed STRAP model using the AUROC, NSS and KL-Div metrics. The advantage of using high-level priors is noticeable. With the three metrics, the inclusion of Gaussian and/or face detection priors always leads to better results in terms of prediction. The presence of a face highly attracts visual attention. For the two sequences with faces, assuming that the face detection performs well, it is not necessary to simultaneously use face detection and center Gaussian priors. On the other hand, in the absence of a face, it is very important to take into account the center Gaussian as high-level information. Therefore, it will be used in the next experiments, but only when compared with models that also integrate a centered Gaussian for the sake of a fair comparison.

7.3.2.2. Validation 2: first viewing salient objects detection

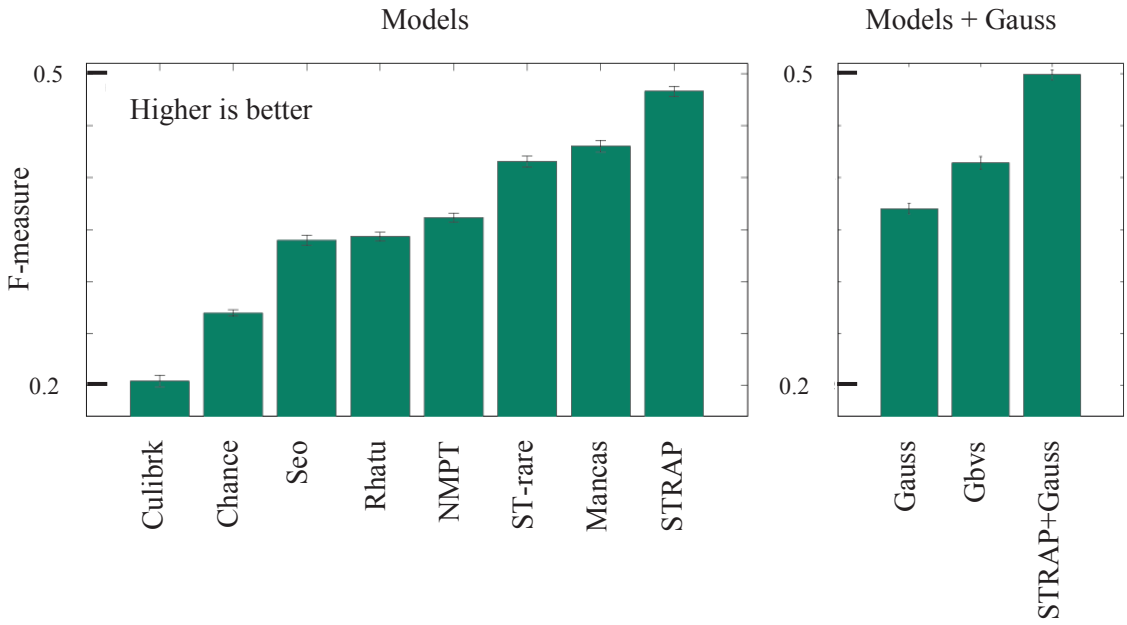


Figure 51 Validation with the binary mask and the F-Measure. (Left) Models comparison without centered Gaussian / (Right) Models comparison with centered Gaussian

To assess the efficiency of the model to detect salient objects, the manual binary masks are used with the F-Measure. Figure 51 shows the performance results. Models that do not use center Gaussian are compared on the left. As a baseline we added the "Chance" model [Judd 2011] that randomly selects pixels as salient, whereas models incorporating center Gaussian are matched up on the right. On the right we added a simple centered Gaussian (called "Gauss") [Judd 2011] as a baseline. We can observe that STRAP achieves the best performance compared with the reference state-of-the-art models, both with and without centered Gaussian.

7.3.2.3. Validation 3: first viewing human prediction

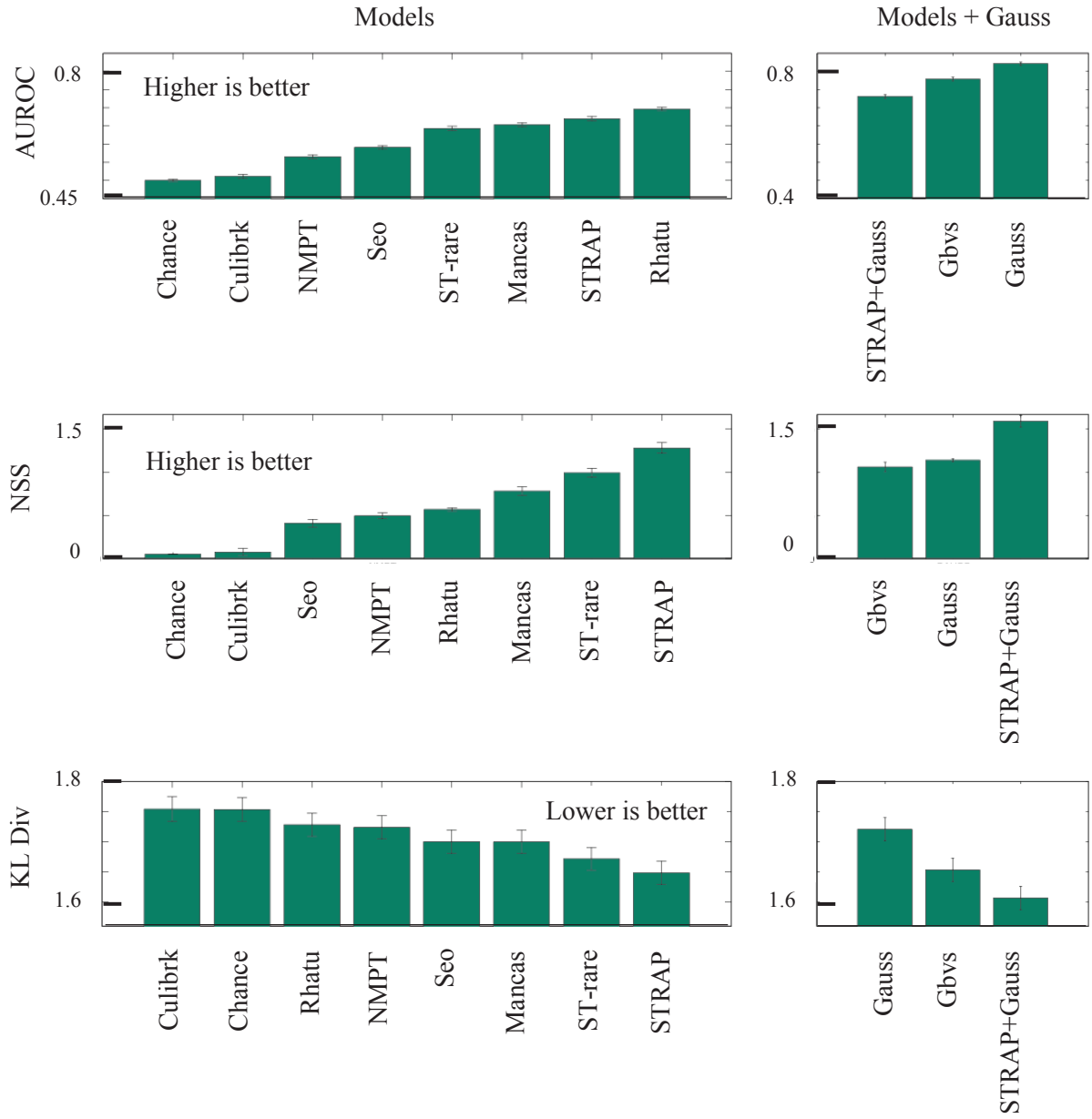


Figure 52 Validation with first viewing human prediction. (Left) Models comparison without centered Gaussian / (Right) Models comparison with centered Gaussian

For this evaluation, the first viewing human prediction is used. In Figure 52, models that do not take into account center Gaussian are compared on the left, and models with center Gaussian are shown on the right. Results for the three complementary metrics AUROC, NSS and KL-Div are given. We can observe that STRAP outperforms all the other models for the NSS and the KL-Div metrics. For the AUROC metric, in the case without centered Gaussian, Rhatu reaches a better score, whereas with centered Gaussian our model is also less good that GBVS. One can also see that both STRAP and GBVS perform less well than the baseline "Gauss" which seems to be the best of all the models with near 0.8. This issue is probably due to the nature of the AUROC metric which only takes into account the location of the salient areas and not their amplitude. In this case models which highlight a lot of the image (like Rhatu 440 and even more like the centered Gaussian) might take unfair advantage. In

order to have a reliable result, one of the three metrics presented here is not enough alone, all the three should be taken into account. When looking to the NSS and KL-Div metrics, STRAP (and G STRAP) for the version which includes the centered Gaussian is clearly the best.

7.3.2.4. Validation 4: a comparison of first and second viewing human prediction

In most evaluations, only the eyes tracking results after a first viewing are used. In this experiment, we would like to evaluate the efficiency of the different models to predict the eyes tracking results after a second viewing. In this case, users have prior knowledge about the scene content. Figure 53 confirms the fact that users do not watch the same parts of the video during the second viewing. Figure 54,, compares the performance results during the first and second viewings. We can observe that for the second viewing, all models perform worse. It is rather predictable as all the models have been developed to characterize eye tracking fixations after a first viewing without any knowledge of the scene. We can observe that the most salient areas are explored during the first viewing, and that the second viewing is essentially used to discover other areas which are less salient. It can also be noticed that the performance drop is less significant for some models only for the NSS metric. Again, one metric alone is not enough to well characterize the models behavior and all the three metric should be taken into account. Nevertheless, the very large drop of the "Gauss" baseline might show that people gaze are less centered and tend to discover more peripheral areas for the second viewing. This study opens new perspective such as trying to model where observers are watching during a second viewing, why some models are more affected than others, and which kind of high priors information can be used to improve the models in this case.

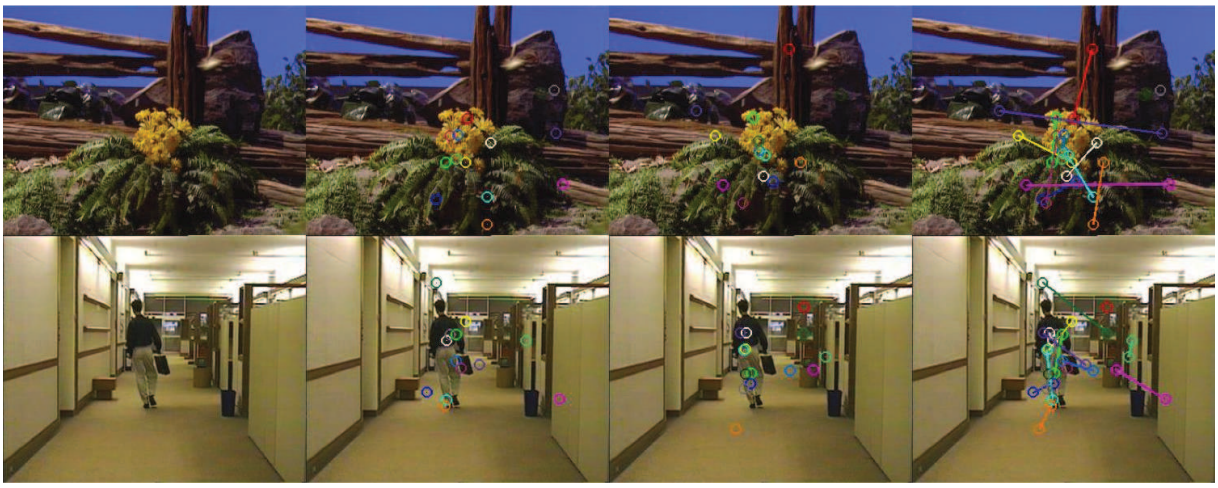


Figure 53 Visual results of the eyes fixations. From the left to the right : Original image, first fixation, second fixation, distance between first and second fixation.

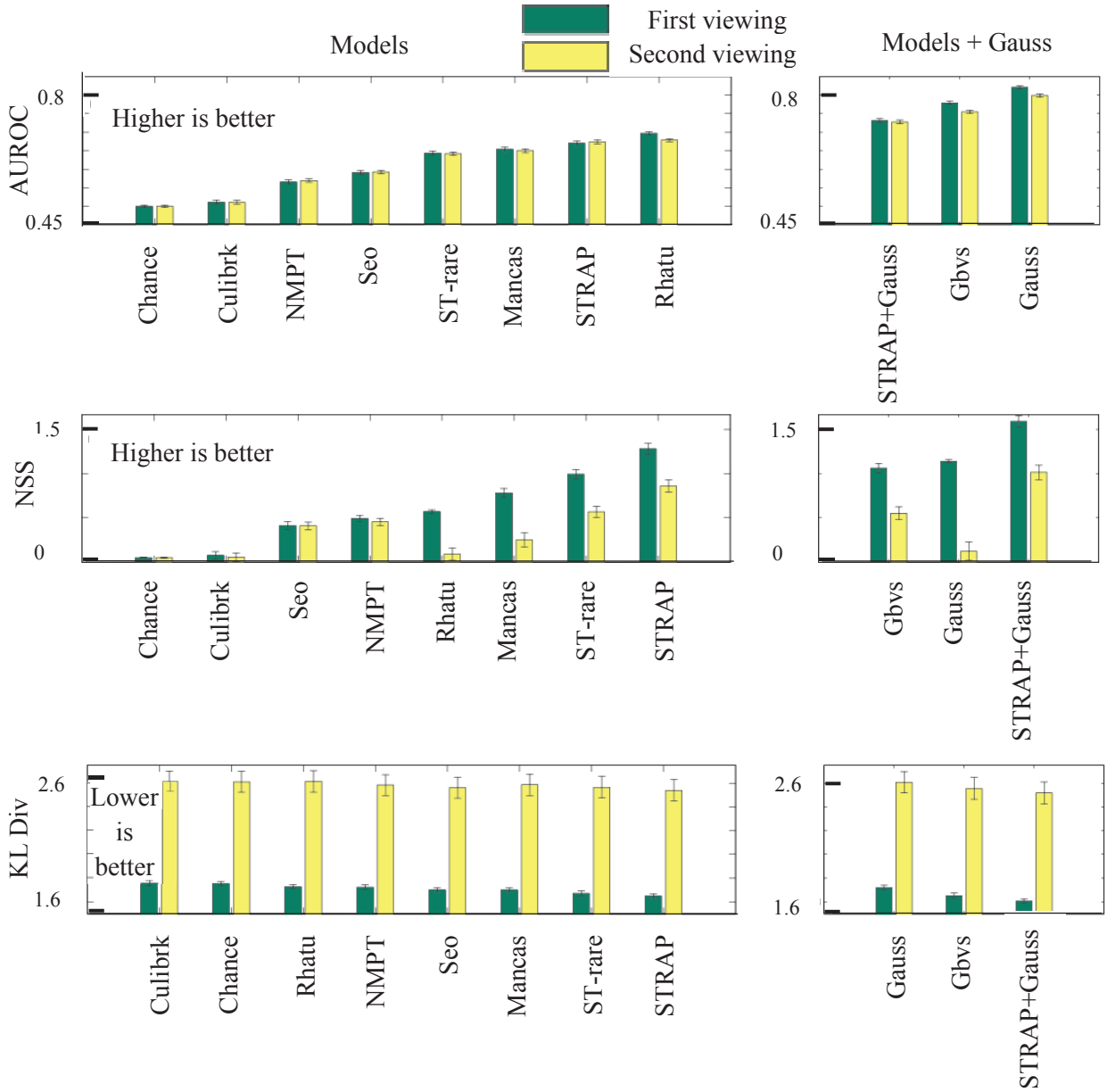


Figure 54 Comparison between first and second viewing

7.4.Results of video compression by seam carving for the [Décombas 2014] approach

After having validated the saliency map, the performance of the video compression based on seam carving has to be assessed. As the approach proposed in [Décombas 2014] is the most recent and advance one, all the evaluation will be performed with this approach.

Figure 55 illustrates the proposed evaluation protocol. After having computed the saliency map as proposed in [Riche 2014] (see Sec. 4.3), the initial seam carving is computed. Then our approach in [Décombas2013a] (See Sec. 5.4) is applied to approximate the seams. The reduced sequence is encoded, along with the modeled seams. To evaluate the results, the reduced sequence is expanded at the decoder side and the final results are compared with the original using the SSIM_SIFT metric [Décombas 2012c] (see Sec. 6.4).

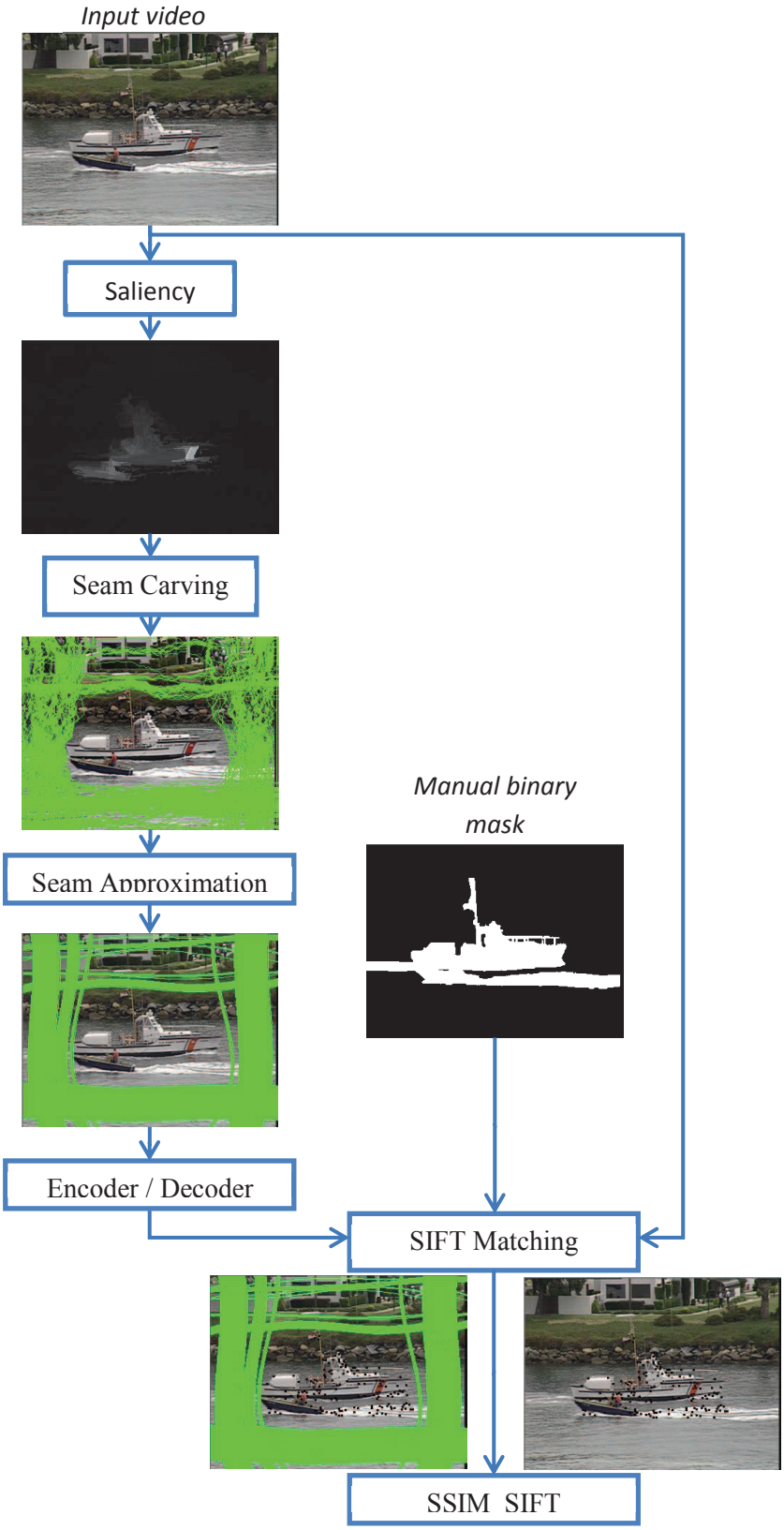


Figure 55 Evaluation protocol. Comparison of the decoded video with seam carving approach with the original video. SSIM SIFT is computed on the black box included in the manual binary mask

7.4.1.Parameters

In our approach, several parameters have been defined. However, their theoretical optimization is quite difficult due to the fact that the proposed approach depends on the video content. Moreover, the evaluation is difficult due to the lack of metrics for some improvements like the seams modeling. The influence of each parameter is described in the Table 6. Extensive tests have been done on different sequences with different ranges of values to heuristically define the optimal value of the parameters.

Parameter	Value	Influence
Saliency map binarisation threshold	$T = 2x\ mean(Saliency)$	Define what is salient in the video and when the seam carving should stop.
Parameter for the content-aware GOP segmentation in chapter 5.4.1	GOP_Threshold =15	Define when a new GOP is created.
Parameters for seams modeling	Spatial_Threshold = 12 Temporal_Threshold = 10	Define the groups of seams
	Outliers_Number_Threshold = 1 Threshold_outliers_Length = Length_Gop/2	Define the outliers

Table 6 Influence and value of the parameters

7.4.2.Rate of spatial reduction

As our approach reduces the spatial dimensions of the sequences depending on the content, the rate of reduction varies accordingly. The rate of reduction is directly linked to two parameters: the binarisation coefficient applied on the saliency maps to obtain the control maps and the content aware GOP cutting. The percentage of pixels removed is defined as:

(34). $$Percentage = 100 * \left(1 - \frac{Spatial\ Dimension\ of\ the\ reduce\ sequence}{Spatial\ Dimension\ of\ the\ original\ sequence}\right).$$

Figure 56 illustrates the evolution of the seams as a function of time. The number of seams that can be suppressed without any constraint is shown in blue, and the proposed approximation in green. A compromise is obtained between a high number of GOPs giving a better approximation, but more subsequences to encode.

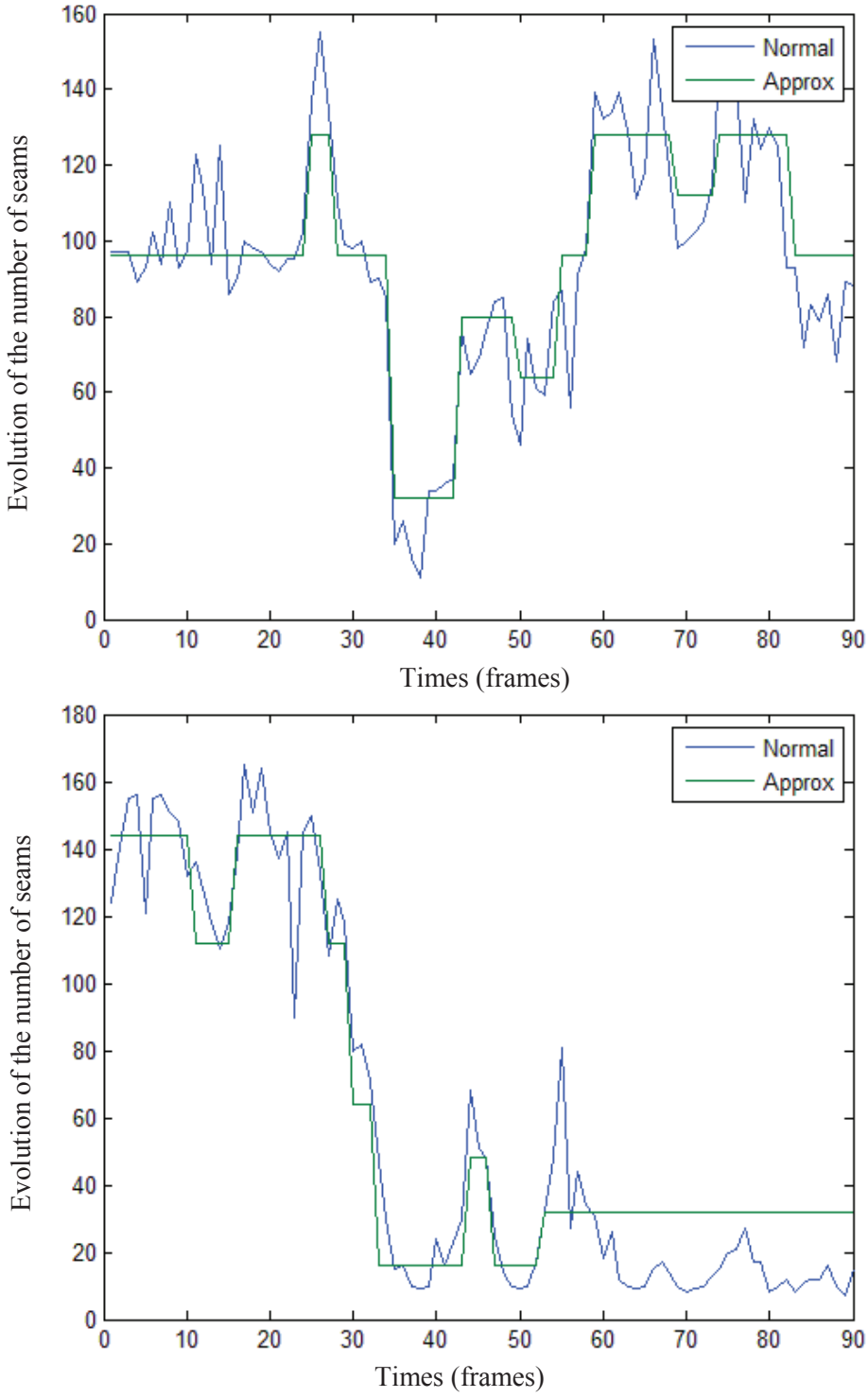


Figure 56 Approximation of the number of suppressed seams for parkjoy. In the top, evolution of the number of vertical seams and in the bottom, evolution of the number of horizontal seams

Sequence	**No constraint**	**Fixed GOP = 5** [Décombas 2011], [Décombas 2012a] and [Décombas 2012b]	**Proposed approach** [Décombas 2014]
Coastguard	53.51	38.45	54.32
Duck	27.69	9.88	26.9
Parkjoy	44.15	30.82	45.42
Parkrun	33.66	17.57	32.64

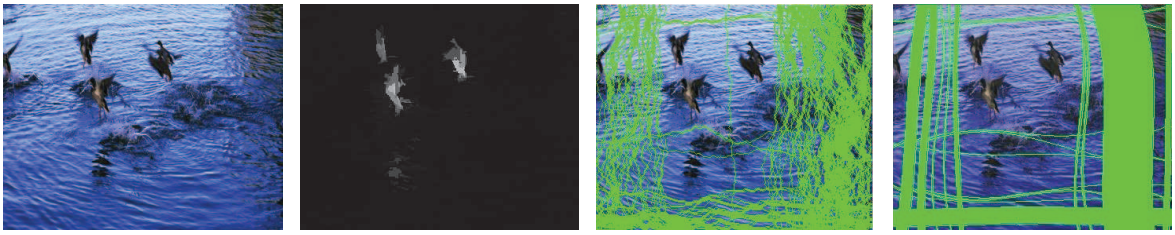
Table 7 Comparison of the rate of reduction between different approaches.

In Table 7, a comparison of the reduction rate between an approach without any constraint on the number of suppressible seams, an approach with fixed length GOP = 5 and our approach is presented. We can see that the proposed approach achieves a very good rate of reduction, more or less equal to an approach without constraint. In contrast, the fixed GOP approach has a much smaller rate of spatial reduction.

7.4.3.Seams approximation

Coastguard

Duck



Parkjoy

Parkrun

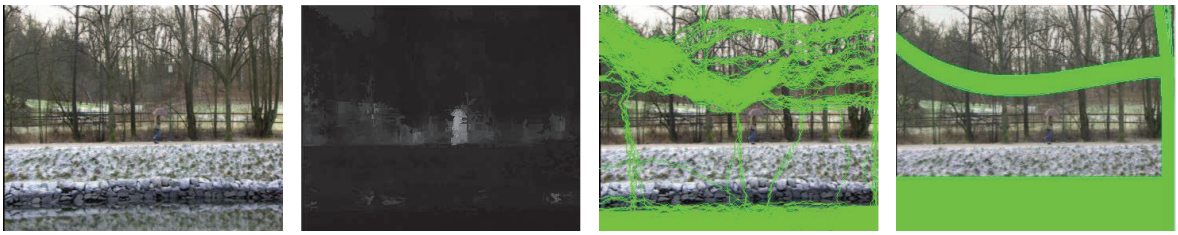


Figure 57 Visual seam modeling for Coastguard / Duck / ParkJoy / ParkRun. In the left, saliency objects from [Riche 2014], in the center original seams and in the right modeled seams.

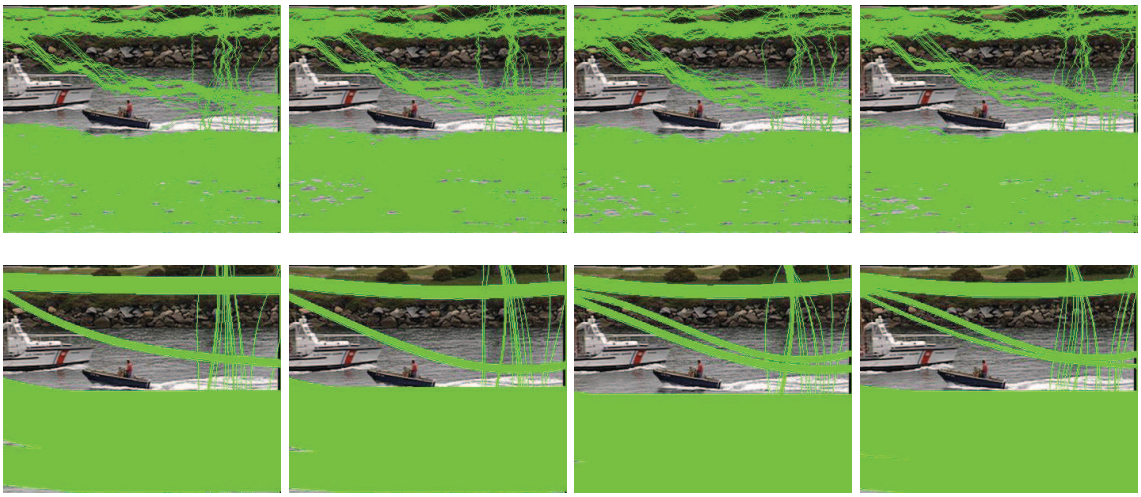


Figure 58 Visual seam modeling for Coastguard at different times. *t* = 27,28,29,30. On first line, initial seams, on second line, the seams after the proposed modeling.

Our approach tries to find a compromise between the flexibility of seam representation and their coding cost. Figure 57 illustrates some visual results for the different test sequences. In general, in all the sequences, the salient objects, obtained using [Riche 2014], are preserved after modeling. Isolated seams are deleted and reallocated to other groups of seams. Some isolated seams are however kept, due to the fact they are consistent in time. Figure 58 illustrates the temporal aspect of the modeling for the Coastguard sequence.

7.4.4. Evaluation of the seam information overhead cost

In this experiment, we evaluate the efficiency of our seams modeling. Three methods are compared:

1. without seam encoding, to illustrate the upper limit bitrate saving which can be achieved due to spatial reduction (note that in this case, the scene geometry cannot be correctly reconstructed at the decoder),
2. our proposed seams modeling and encoding
3. all seams positions are fully encoded without seam modeling.

H.264/AVC is used in full intra coding that is consistent with video surveillance applications. The bitrate of the entire sequence is taken into account during the comparison. The Quantization Parameter Intra (QPI) varies from 27 to 39 with a step of 3. Table 8 illustrates the percentage of bitrate saved compared to H.264/AVC as a function of QPI.

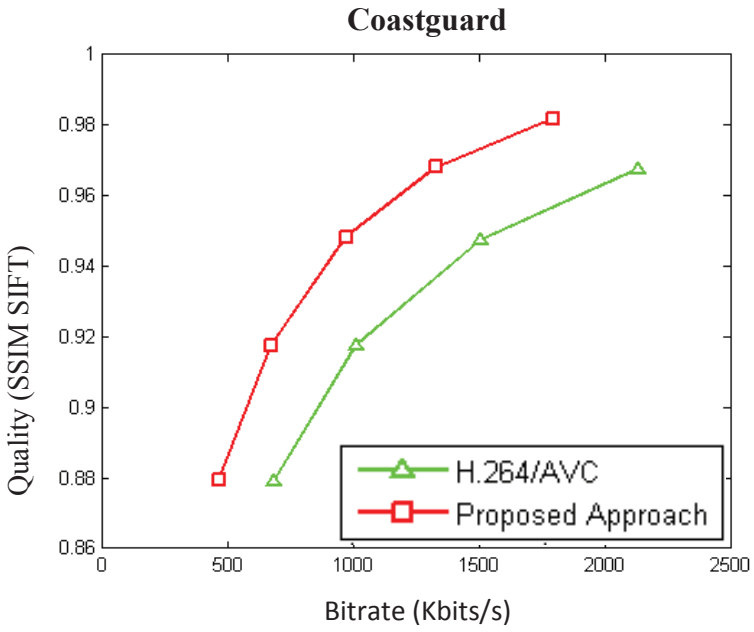
	coastguard			Duck			ParkJoy			Parkrun		
Rate of reduction	**54.32 %**			**26.9 %**			**45.42 %**			**32.64 %**		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
QPI=27	39.78	39.29	-1.025	24.27	23.94	-6.46	26.75	26.47	-11.06	27.69	27.44	15.02
QPI=30	37.51	36.82	-19.62	23.55	23.10	-18.25	24.70	24.33	-26.73	27.08	26.76	10.20
QPI=33	35.61	34.64	-45.08	22.43	21.79	-36.74	22.90	22.36	-50.87	26.54	26.09	3.19
QPI=36	33.42	31.98	-87.00	21.69	20.74	-66.47	21.49	20.66	-91.89	20.02	25.34	-8.60
QPI=39	31.56	29.42	-146.5	21.13	19.72	-111.4	20.71	19.42	-155.4	25.61	24.59	-26.57

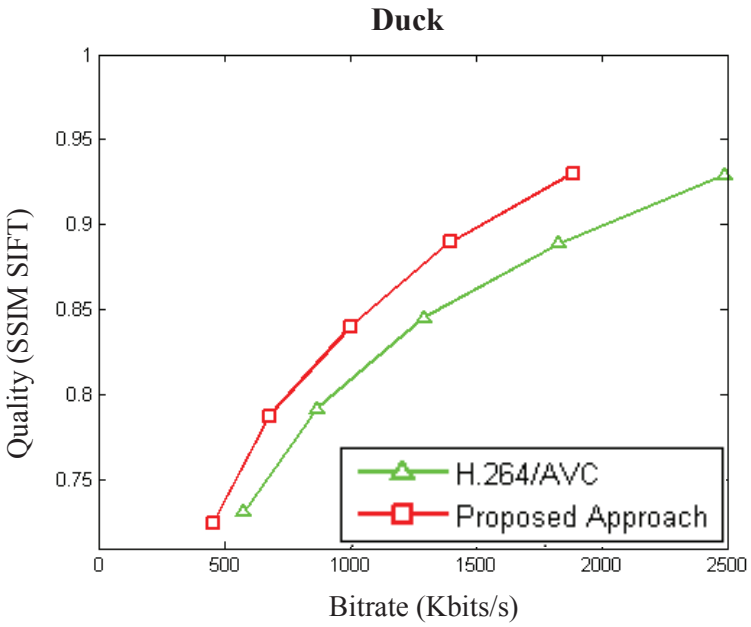
Table 8 Percentage of bitrate saved compared to H.264/AVC as a function of QPI (positive value means bitrate is decreased, negative value means bitrate is increase): (1) Without seam encoding, (2) Our proposed approach, (3) Total seam position coding without modeling

We can see that when all the seam positions are encoded, the overhead cost becomes too important and a classical video encoding is better. This justifies the fact that the seams position has to be approximated. At the opposite, no seam encoding approach reaches the optimal bitrate saving for a given spatial reduction. However, as the seam carving is not a reversible process, artifacts may appear at the decoder in this case. In the proposed approach, we can see that the overhead due to the seam encoding is low and the saved bitrate is very close to the approach where no seam is encoded.

7.4.5. Rate distortion performance assessment

Finally, we assess the rate-distortion performance of the proposed semantic video coding scheme based on seam carving. For this purpose, comparison is made in full intra coding, which is very common and pertinent in security applications. The traditional H264/AVC encoder is used as reference. The SSIM_SIFT metric in [Décombas 2012c] is used to measure quality. This metric requires a binary mask to define the salient objects. In these experiments, we have used the manual binary masks from [Riche 2014]. It should be underlined that these binary masks are only used to compute SSIM_SIFT.


Coastguard
Quality (SSIM SIFT)
Bitrate (Kbits/s)
H.264/AVC
Proposed Approach
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0
500
1000
1500
2000
2500


Duck
Quality (SSIM SIFT)
Bitrate (Kbits/s)
H.264/AVC
Proposed Approach
1
0.95
0.9
0.85
0.8
0.75
0
500
1000
1500
2000
2500

.

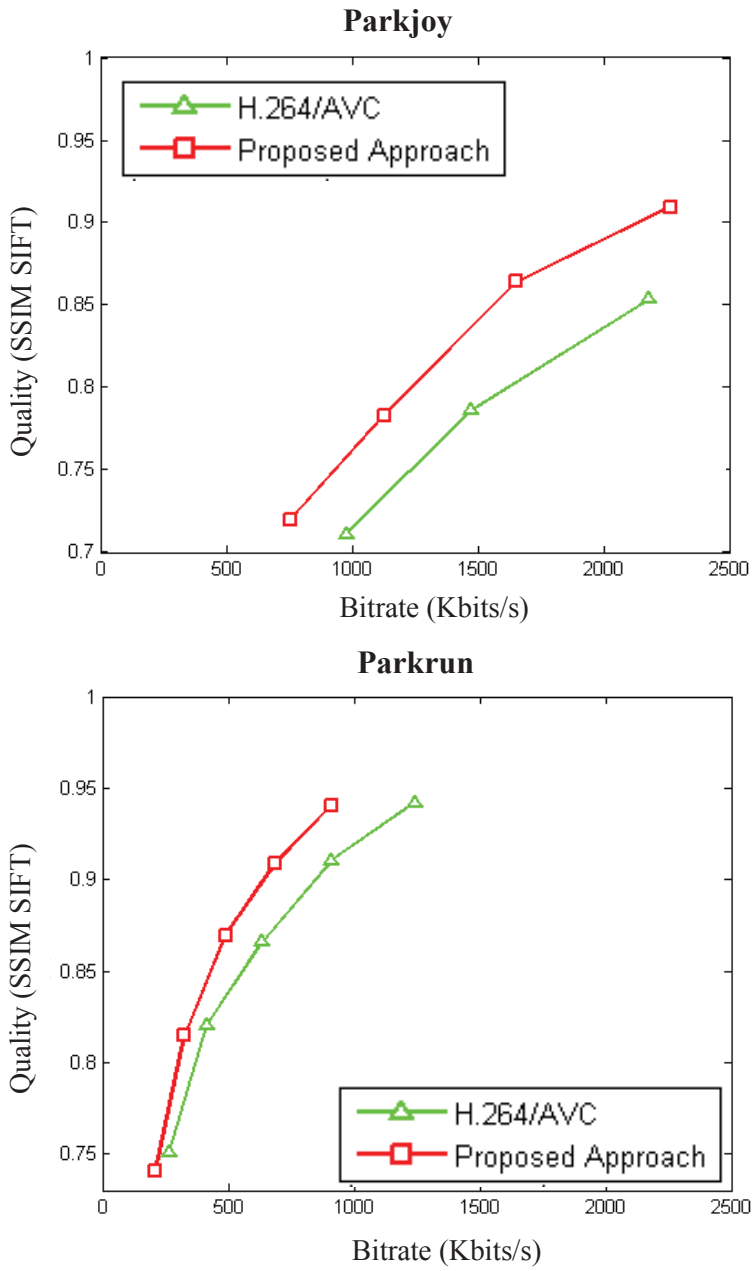


Figure 59 RD curves for Coastguard, Ducks, Parkrun, Parkjoy. In green, H.264/AVC and in red the proposed approach.

The proposed approach consistently outperforms H.264/AVC for all test sequences. The Bjontegaard's metric [Bjontegarrd 2001], used with the SSIM_SIFT [Décombas 2012c], allows to compute the average percentage of bitrate saving between the two rate-distortion curves. The bitrate saving achieved by the proposed scheme is 37.27% for Coastguard, 19.07% for Duck, 26.45% for Parkjoy, and 18.30% for Parkrun.

Compressed video with seam carving (789 Kbits/s)

H.264 /AVC (985 Kbits/s)

Figure 60 Rate distortion performance. Visual results for Parkjoy with a SSIM_SIFT = 0,72 and a bitrate saved of 19,4%

7.5.Conclusion

In this chapter, raw video database where eyes tracking and manual binary mask are available is proposed [Riche 2014]. So, evaluation for the saliency model [Riche 2014] (see Sec.4.3) has been realized with three different metrics on the eyes tracking reference and one metric for the manual binary mask reference. Four validations have been done. The first study measures the usefulness of high levels priors, the second study evaluates the first viewing salient object detection with the manual binary mask, the third study evaluates the first viewing human prediction. The last study is a comparison between a first and a second viewing human prediction. It has been seen that our models outperforms in general all the others models. The last study shows that all the models give bad results after a second viewing. In the case of video compression, the models presented in [Décombas2013a] (See Sec. 5.4) is evaluate. First the choice of the parameters has been justified. Then some results about the rate of spatial reduction and the seams approximation are shown. An evaluation of the seam information overhead cost is realized and rate distortion curves are plotted using the SSIM_SIFT metric [Décombas 2012c] (see Sec. 6.4).

8. Seam carving for video summary

8.1. Introduction

In the previous chapters, we have seen the advantage to use seam carving to suppress non salient information and in this way, reduce the quantity of information to send. The problem of too large volumes of information to manage is well known in video surveillance.

In the last years, millions of video cameras have been deployed in the city's streets, the highways and transportation hubs. In 2007, the numbers of surveillance cameras have reached 30 millions in the United States, producing over four billion of video footage each week [Vlahos 2008]. This proliferation of cameras is due to an inexpensive camera network, easily deployed and remotely managed. Therefore, the development of automatic algorithms to aid human operators to identify what is important and store it in a video summary becomes essential.

In this chapter, we present algorithms that compute a video summary for video surveillance applications, as proposed in [Décombas 2013b] (8) and [Décombas 2013c] (9). The video summary is an abbreviated video that preserves salient parts while removing spatio-temporal segments without interest. Such a summary is especially useful in video forensics, but can also be applied to quick video review for home applications [Oh 2004]. First, existing methods and their limitations will be reviewed, then the proposed method will be presented and visual results will be discussed.

We propose to perform video summary using seam carving with a spatio-temporal grouping constraint. Seam carving is applied in the (x,t) plane, with the aim of a reduction along the temporal dimension. Our approach first computes all the seams and then analyzes their evolution in space and time. We propose (8,9 a) a way to do an efficient spatio-temporal grouping that allow us (8,9 b) to determine a temporal rate of reduction in function of the content, (8,9 c) to suppress the group of isolated seams, (8,9 d) to identify sufficiently large spatio-temporal groups of seams, and (8,9 e) to approximate by constant segments the number of seams for each group, while keeping the total sum of seams constant. Applying the seam carving directly for video summary leads to geometric deformations of the salient objects, anachronisms and a summary without the same length on all the lines. This is due to the fact that the quantity of information suppressed between the salient objects is not the controlled. The proposed method limits the artifact and gives a summary with the same length on all the lines. In addition, the summary has the same length on all the lines. Our constraint enables more flexibility in the seam carving process, with seams that can better adapt to the content. At the same time, a better rate of temporal reduction can be achieved while preserving salient objects.

8.2. State of the art

Three main approaches to perform a video summarization have been proposed.

8.2.1. Fast forwarding

Firstly, with fast forwarding, frames are skipped in fixed or adaptive intervals. In the case of fixed intervals, salient objects that do not stay long enough are not visible. The approaches in [Yeung 1997], [Nam 1999], [Petrovic 2005]) propose to skip frames at adaptive intervals. However, the limitation of these techniques is that, as only complete frames can be removed, the rate of temporal reduction, defined as the ratio of lengths of the original and processed videos, is relatively low.

8.2.2. Key frames extractions

A second approach to video summarization is to extract key frames and present them simultaneously as a storyline, as reviewed in [Oh 2004]. Due to the fact that the keys frames can be very far away from each other, the temporal continuity of events is not preserved. Therefore, important contextual information is lost. An extension of this approach is to extract short sub-sequences, but the problem of having a relatively low reduction rate remains.

8.2.3. Spatio-temporal combination

The two previous approaches preserve complete frames. The third approach is to extract salient spatio-temporal segments and combine them to obtain a video summary. In [Irani 1996], Irani *et al.* combines spatial segments from different times to obtain a single image. A very high condensation of the content is obtained without losing any information but the dynamic aspects are lost. Based on the same idea, spatial segments can be shifted in time to obtain a summary. Kang *et al.* proposes in [Kang 2006] to do video montage by modifying the location of spatial segments. This leads to a good rate of temporal reduction, but also to visible artifacts due to the combination of uncorrelated segments. In [Rav-Acha 2007], a solution is proposed for panning camera where segments are aligned temporally and viewed simultaneously. Another approach of this work is proposed in [Rav-Acha 2007], [Pritch 2008], [Pritch 2009], where an alternative to video montage called video synopsis is presented. Dynamic objects are identified, extracted and combined by using the minimization of a cost function. To prevent the total loss of context that can appear when spatio-temporal shifting is allowed, only temporal shift is allowed. The approach performs well, but the displacement of segments may cause a reversal of the order of activities. Another limitation of this approach is its complexity, as it involves several computationally complex tools such as object detection and background subtraction, clustering, and combination of spatio-temporal segments.

To cope with this puzzle, approaches have been recently proposed based on seam carving [Avidan 2007]. In [Rubinstein 2008], Rubinstein extends seam carving to video applications and proposes a new way to compute the seams that take into account the influence of their suppression. A direct application of the seam carving without any constraint for producing the video summary creates temporal anachronisms, deformations of the objects and a summary not having the same length on all the lines. In [Chen 2008], [Li 2009] the authors propose to compute seams one by one with a spatio-temporal constraint. More specifically, the spatial constraint is the flexibility of the seam to move by one pixel to the left or to the right during its computation. The temporal constraint is the capacity of the spatial seam to evolve from one frame to the next one.

Chao et *al.* propose in [Chao 2011] to compute a seam at frame *t*-1 and use the black based motion estimation and Gaussian masks to predict the coarse location of the seam in the frame *t* that reduce the search range of dynamic programming and allows having seams that can move with more than 1 pixel from one frame to another.

Ishwar and Konrad propose also a solution to the problem of seams flexibility following the temporal axis [Li 2009]. They introduce the concept of ribbon and carve ribbons out by minimizing an activity-aware cost function using dynamic programming. Their model permits an adjustment of the compromise between temporal condensation ratio and anachronism of events. The problem of GOP is solved by using a sliding-window ribbon carving. By this way, they can handle streaming video.

Therefore, we propose to first compute all the seams in the (*x*,*t*) plane for all the *y*. Then, we apply our spatio-temporal grouping constraint.

8.3.Temporal seam carving with groups

8.3.1. General approach

Considering a video as a cube of dimensions (*N*,*M*,*T*), the seam carving can be applied in the (*x*,*t*) plane with a reduction along the temporal axis, thus leading to a summarized video. Figure 61 presents the proposed approach of video summarization using seam carving with spatio-temporal grouping constraint. From an original video of length *T*, saliency maps are created with the ST-RARE model [Décombas 2013a] computed in the (*x*,*y*) plane.

This model identifies the salient objects by finding the rarity on different maps. The most pertinent maps are combined together to obtain a unique saliency map. The original model uses static (L,a,b) and dynamic (amplitude and direction) components in order to identify salient areas for static and moving scenes. However, in the absence of movement, the model identifies static salient areas, preventing the suppression of the otherwise insignificant frames. Therefore, in our approach, only the dynamic component of the model is taken into account for video summarization.

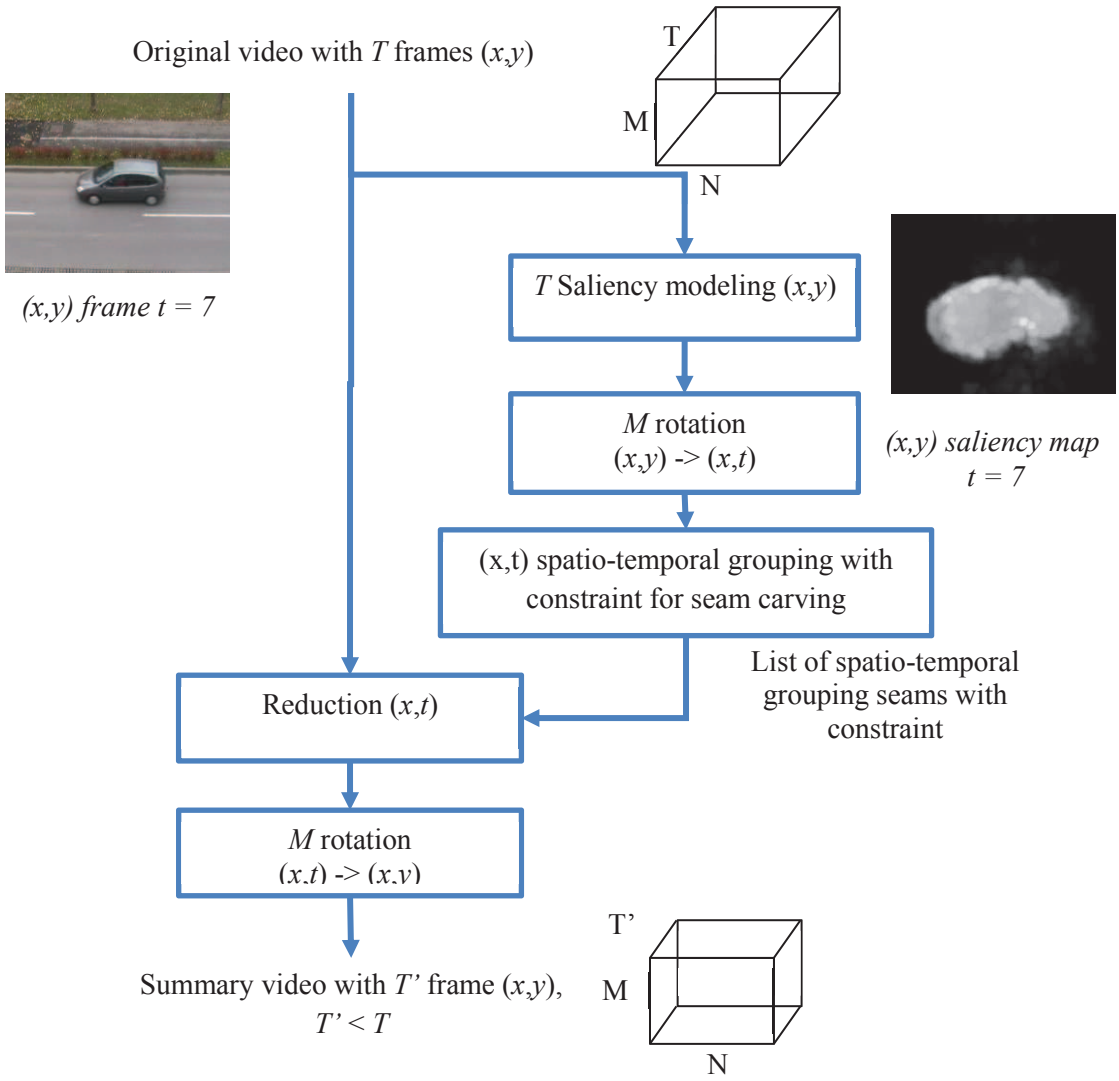


Figure 61 Proposed approach of video summary with spatio-temporal grouping constraint for seam carving.

To compute the seam, the forward energy of Rubinstein [Rubinstein 2008] is used on each frame. Contrary to [Rubinstein 2008] which uses a graph to define the seam and then force the seams to have a temporal link, dynamic programming is used on each frame separately in our approach. The seam carving with spatio-temporal grouping constraint is then applied on the (*x*,*t*) plane, giving a list of seams in each frame (*x*,*t*). These seams allowing to reduce *T* are suppressed in the original video to obtain a summarized version of length *T'* < *T*.

8.3.1.1. Rate of temporal reduction

Seam carving is an iterative process allowing to pass from a resolution to another one depending on a stopping criteria. In most applications, seam carving is used to change the spatial resolution and the target output resolution is known a priori. In these cases, the stopping criterion is a vertical and horizontal number of seams to suppress.

In our case, the objective is to reduce as much as possible the temporal aspect while preserving the salient objects. For this purpose, a binarisation of the saliency map is used to separate the salient content that has to be preserved from the rest of the video. This binary saliency map is used as stopping criterion. While a seam does not cross the binary saliency map, the frame (*x*,*t*) passes to (*x*,*t*-1). Seam carving is applied independently on each frame (*x*,*t*) and stopped in function of the motion activity. As the objective is to preserve all the salient areas, and as some frames have more activity than others, the rate of temporal reduction is defined as the minimal number of suppressible seams. This step is the block "Uniformization of the number of seams" in Figure 62.

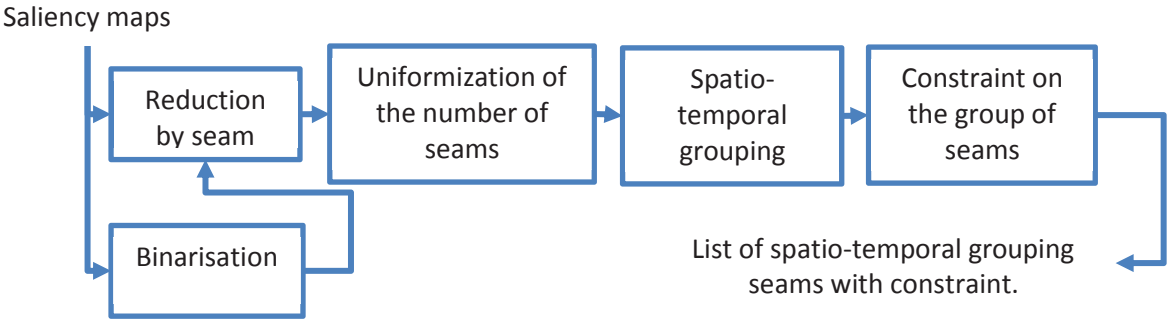


Figure 62 Spatio-temporal grouping with constraint for seam carving.

8.3.1.2. Spatio-temporal grouping

If the seam carving is applied without any spatio-temporal constraint, some artifacts may appear as it can be observed in the (c) image of the Figure 67, Figure 68, Figure 69 and Figure 70 . The proposed approach reduces this problem by creating groups of seams. By this way, the salient objects will shift by the same amount in time.

To perform the spatio-temporal grouping, the seams are first grouped together with a *Spatial_Distance* and a *Spatial_Threshold.* The *Spatial_Distance* is defined as

(35). $D(Seam_t, Seam_{t+1}) = \max_{i=1\ldots N}\left(Seam_t(i), Seam_{t+1}(i)\right)$

where *N* is the length of a seam. This distance has been chosen as it successfully identifies salient objects between seams, contrary to the mean or the median. The Spatial_Threshold represents the maximal distance between two consecutives seams found in the same group. It has been experimentally defined at 7 pixels. Then, the groups have to be temporally linked together. For this purpose, the symmetric difference is used between the groups of seams areas at *t* and at *t*+1. The symmetric difference is defined as the area of the union minus the area of the intersection. The area of the groups of seams is bordered by the most left seams and the most right seams. Then, the temporal regrouping is done in function of the minimum of this difference. These steps are included in the block "spatio-temporal grouping" of Figure 62.

8.3.1.3. Constraint on the group of seams

Next, the groups of seams are processed as illustrated in Figure 63 to solve the problem of geometric distortions and anachronism. The number of seams by group varies with the time ($Nb\_Seam_{Gpe\_x}=f_{Gpe\_x}(t)$). Isolated groups, not present in enough frames, are suppressed. The associated number of seams is reallocated to other groups. Then, a median filter is applied temporally on each function f_{Gpe_x} to suppress strong local variations. To avoid anachronism, the function f_{Gpe_x} has to be piece-wise constant. The length of the pieces is linked to the size of the salient objects and the time they remain in the scene. Rupture detection is done to segment the function in pieces. The rupture detection is based on the variation of the function compared to the median. The median value is associated to each piece. A set of constant pieces is then obtained for each group.

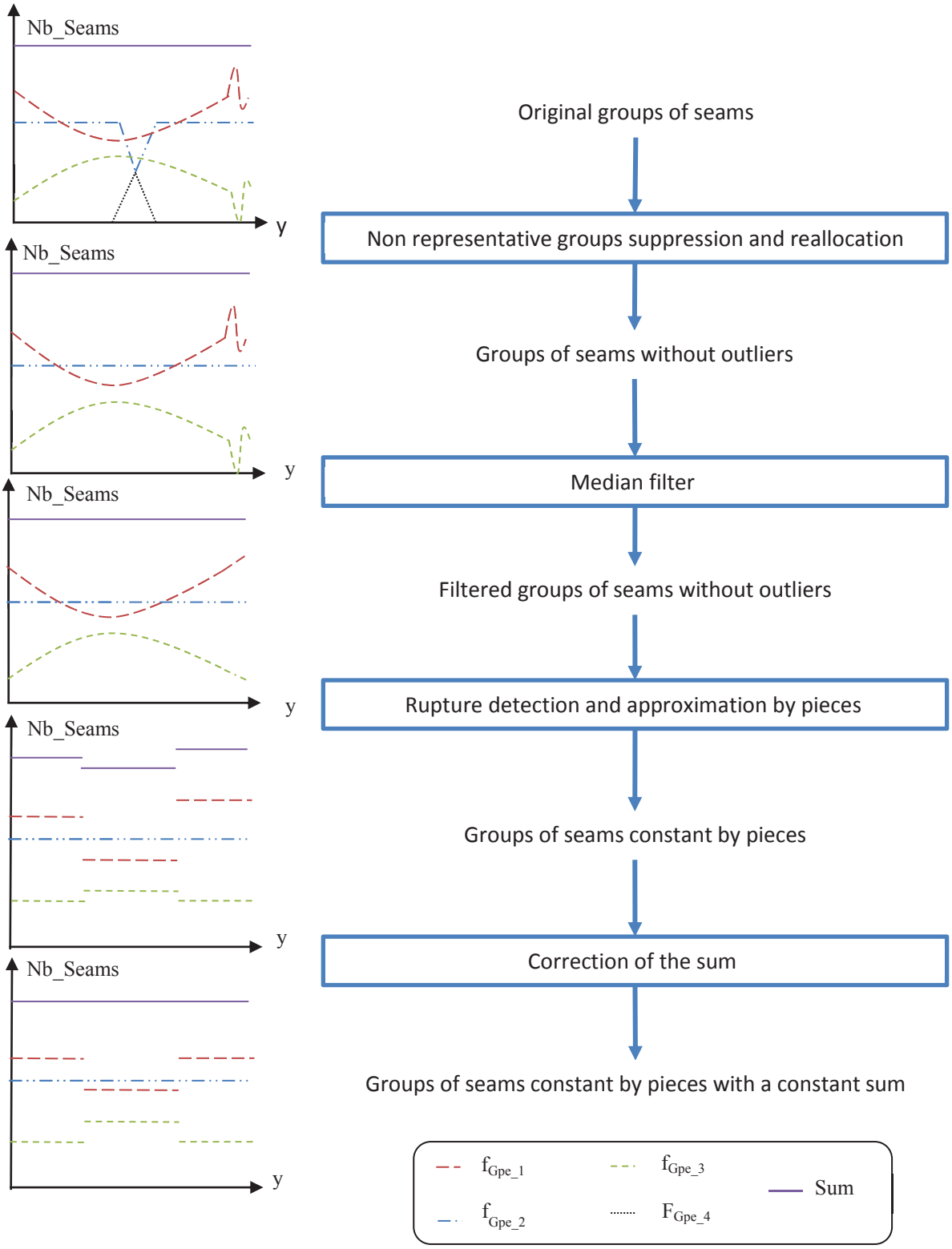


Figure 63 Processing applied on group of seams. From original groups to piece-wise constant groups.

8.4. Experimental Results

To evaluate the proposed method, video surveillance test sequences have been chosen with a fixed camera and salient objects (cars, bicycles) crossing the scene from left to right or from right to left. In

addition, many frames do not contain any activity. This type of sequences is very representative of video surveillance applications.

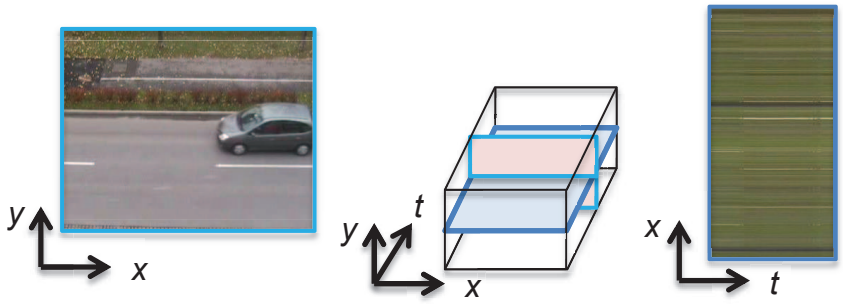
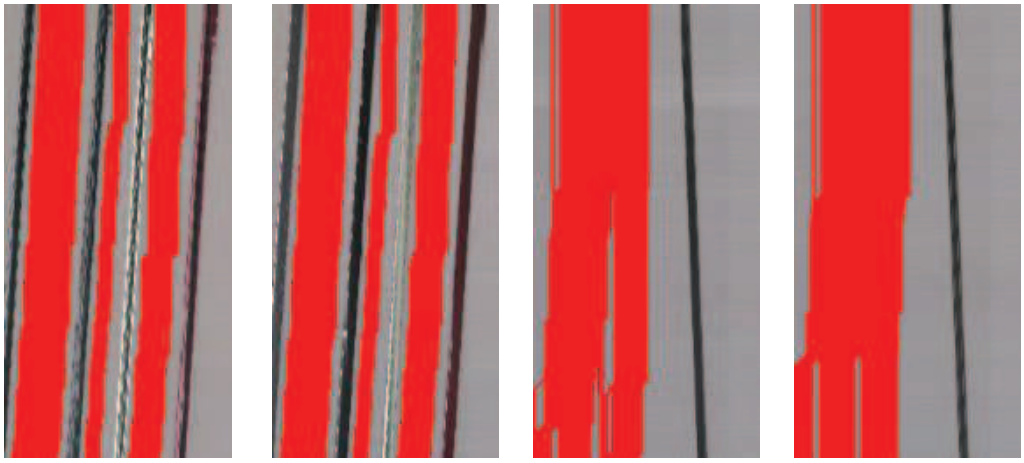


Figure 64 Visualization of original image in the plan (*x*,*y*) and (*x*,*t*)

Figure 64 illustrates the original video on the plan (*x*,*y*) and the plan (*x*,*t*). To do spatial resizing, the seam are computed in the plan (*x*,*y*). As our objective is to do video summary, the seam carving is computed on the (*x*,*t*) plan.

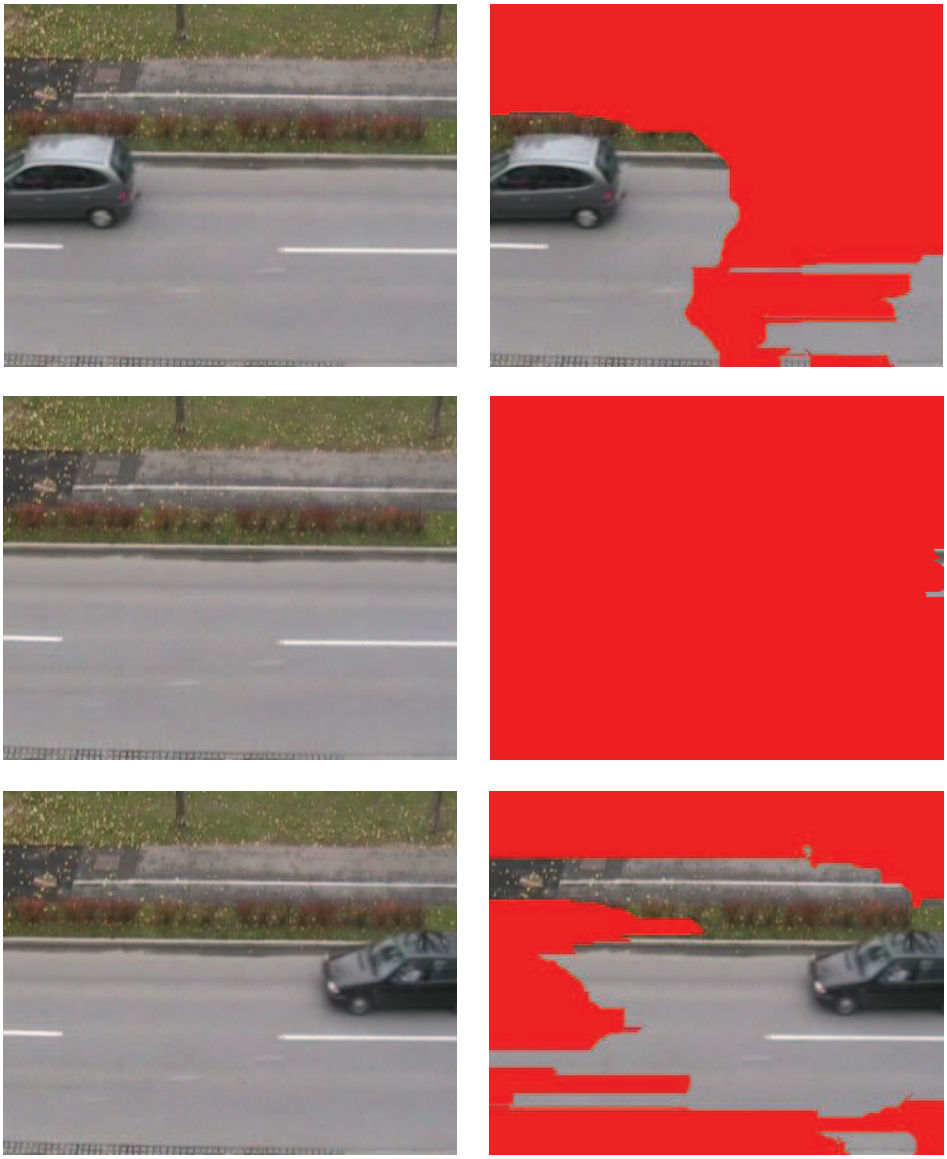
In Figure 65, the suppressible seams (in red) can be seen after the permutation for *y* = {139,154, 244, 253}. The road is in grey and the vehicle trajectories are in black or white depending on the vehicles colors. On the images (a) and (b), the trajectories of 4 vehicles going from left to right on the top of the road are visible. On the images (c) and (d), the trajectory of one vehicle going from right to left on the bottom of the road is visible. With our approach, the quantity of seams suppressed between the vehicles is constant, as it can be seen between *y* = 139 and *y* = 154 or *y* = 244 and *y* = 253. The seams can adapt to the number of salient objects and their trajectories, as it can be seen on all the frames (*x*,*t*) of Figure 65.

The suppressible seams from the Figure 65 on the (*x*,*t*) plane for *y* = {139,154,244,253} can also be seen in Figure 66 for *t* = {17,40,58}. The seams avoid the salient objects at *t*= {17, 58} and when there is no activity in the frame, the frame is totally suppressed, as at *t* = 40.



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

Figure 65 Visualization of the *(x,t)* plane for video-A (a) *y*=139, (b) *y*=154, (c) *y*=244, (d) *y*=253; the suppressible seams are in red, the road is in grey, and the different vehicle trajectories are in black or white.



(a) (b)

Figure 66 Visualization of the original frame for video-A (a) at *t* = 17, 40, 58 and the frame with the suppressible seams (b) at *t* = 17,40,58

Results of our approach are shown in Figure 67, Figure 68, Figure 69, Figure 70. In Figure 67 (a) and Figure 67 (b), vehicles from the original video at *t* = 17 and *t* = 58 are shown. On Figure 67 (c) the approach to do temporal summary based on seam carving without constraint is illustrated. The second vehicle arrives earlier but some geometrical artifacts have appeared. This is due to the fact that the number of seams suppressed before and after the vehicle is not constant on all the lines. As a consequence, pieces of the vehicle arrive earlier than the rest of the vehicle. In our approach, all the lines of the second vehicle are shifted by the same quantity. The vehicle has no artifacts, as can be seen in Figure 67 (d). In Figure 68 (a), (b), two vehicles are visible at *t* = 77 and *t* = 94 in the original video. In Figure 68(c), the approach without constraint deforms the shape of the vehicles, which now

appear shifted in time at t = 38. With our approach in Figure 68(d), the same temporal shifting is obtained but the vehicles are well preserved.

Similar results are shown in Figure 69 and Figure 70. Artifacts are visible on the road in Figure 69 due to a temporally changing luminance. In all cases, the seam carving without constraint is clearly less efficient than the proposed approach. The video-A has 35 frames without activity and the process allows to suppress 68 frames on the total sequence of 180 frames. The video-B has 60 frames without activity and the process allows to suppress 77 frames on the total sequence of 140 frames. Finally, the video-C has 52 frames without activity and the process allows us to suppress 66 frames on the total sequence of 120 frames.

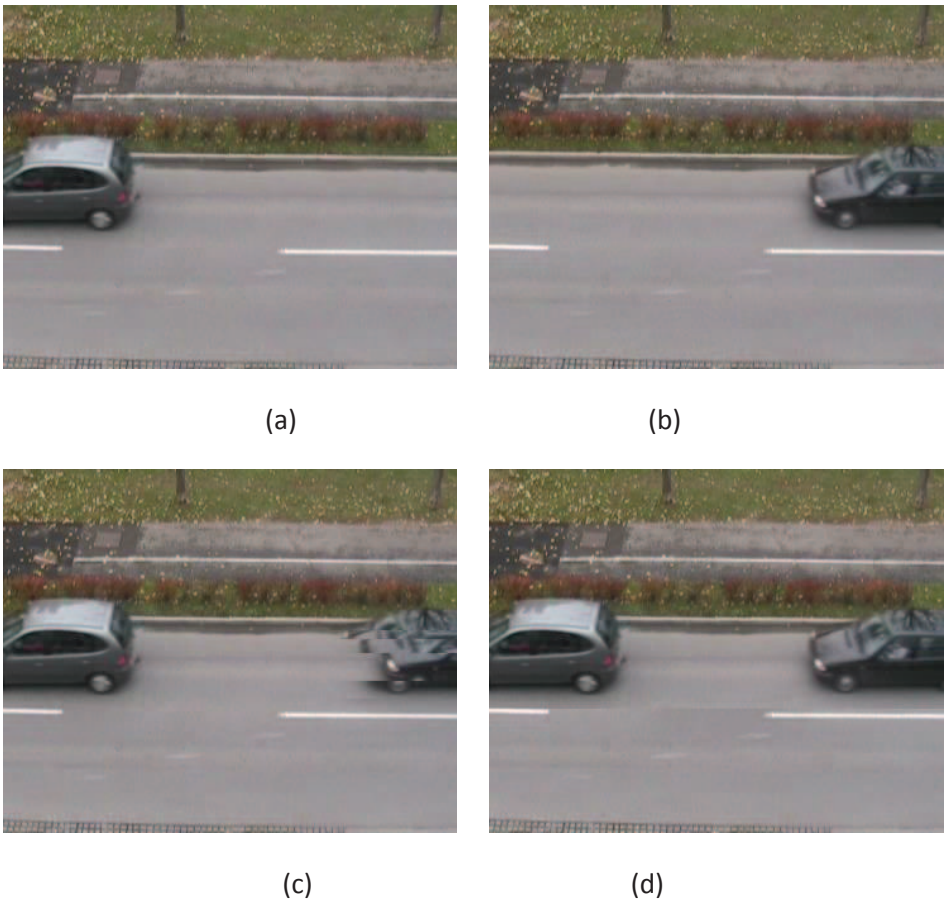
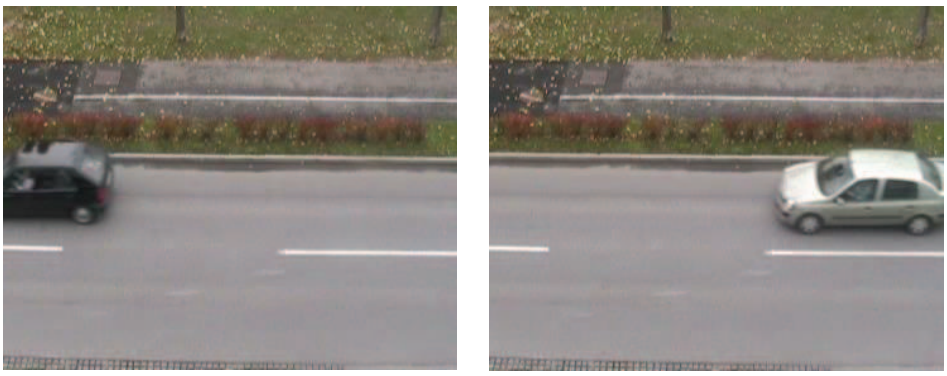
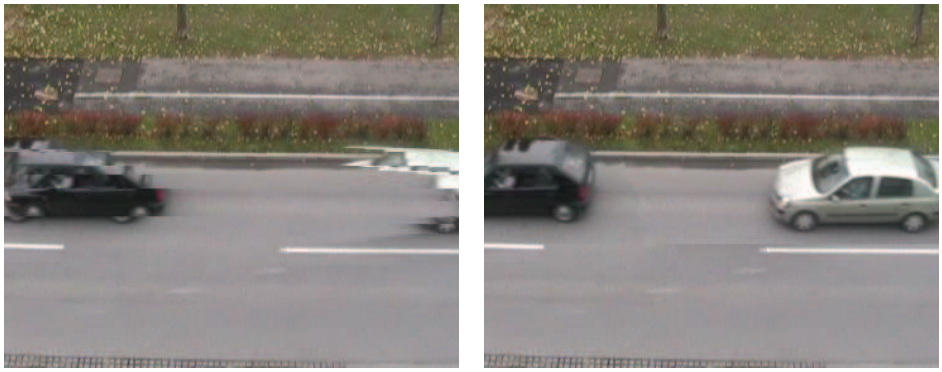


Figure 67 Video-A summary: (a) original frame at t = 17, (b) original frame at t = 58, (c) frame after seam carving without constraint at t= 17, (d) frame after spatio-temporal grouping with constraint for seam carving at t=17

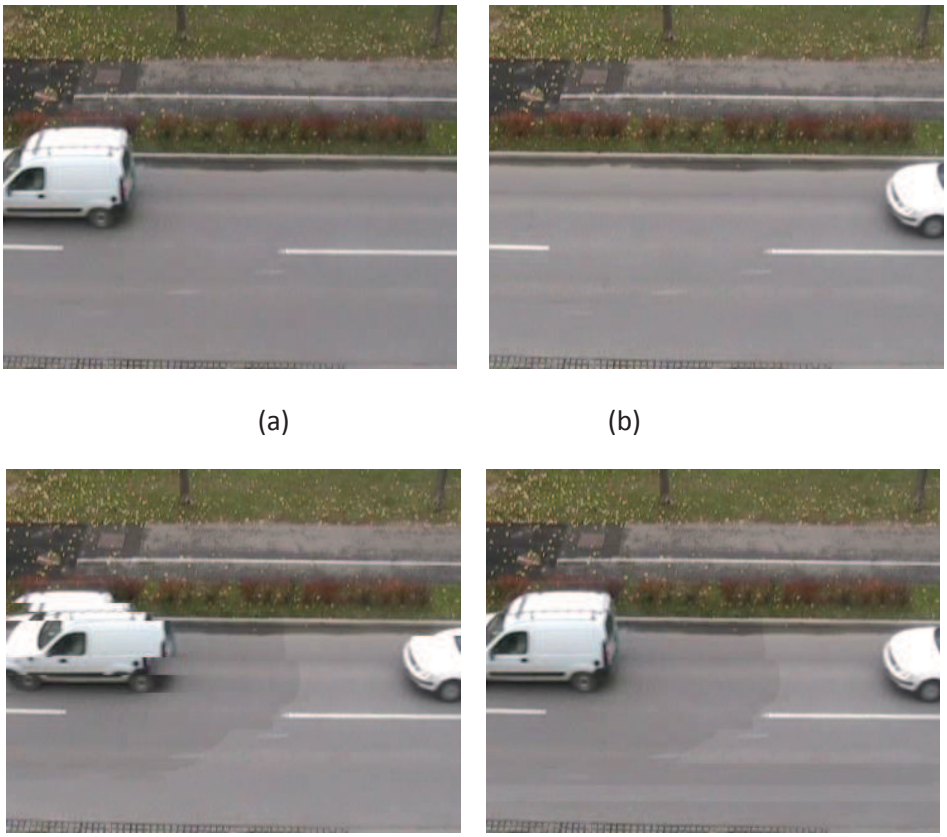


(a) (b)



(c) (d)

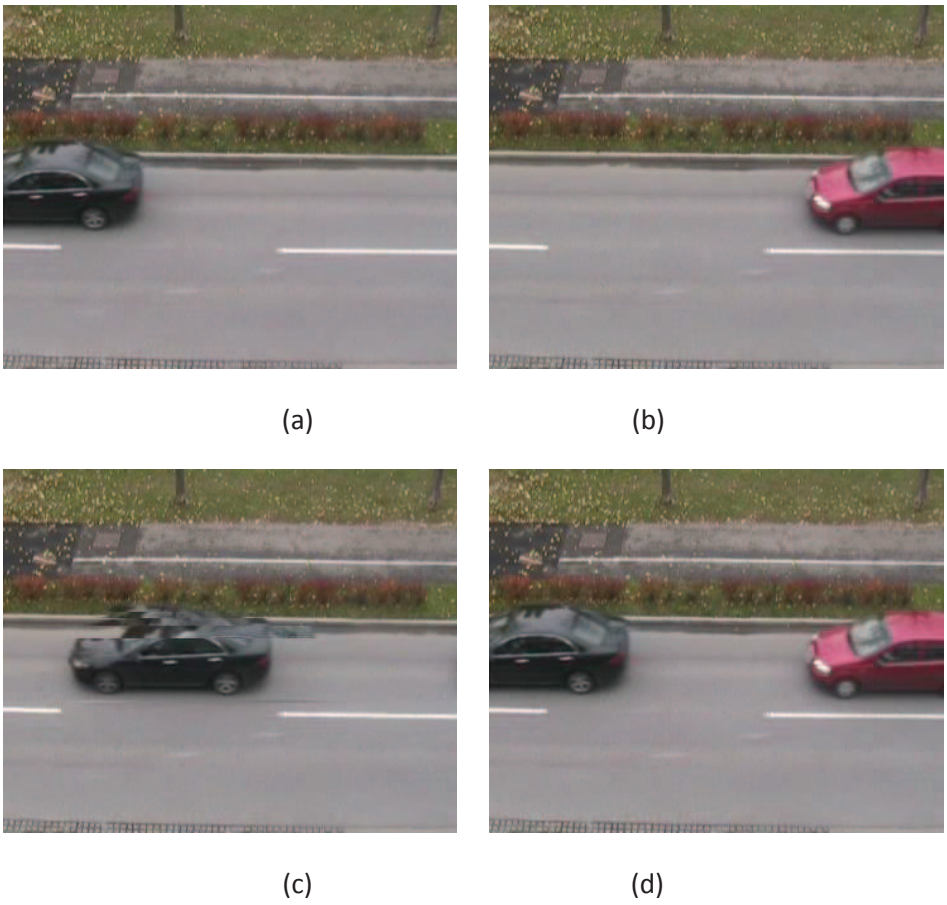
Figure 68 Video-A summary: (a) original frame at t = 77, (b) original frame at t = 94, (c) frame after seam carving without constraint at t = 38, (d) frame after spatio-temporal grouping with constraint for seam carving at t = 38



(a) (b)

(c) (d)

Figure 69 Video-B summary: (a) original frame at t = 34, (b) original frame at t = 106, (c) frame after seam carving without constraint at t = 29, (d) frame after spatio-temporal grouping with constraint for seam carving at t = 29.



(a) (b)

(c) (d)

Figure 70 Video-C summary: (a) original frame at *t* = 29, (b) original frame at *t* = 85, (c) frame after seam carving without constraint at *t* = 20, (d) frame after spatio-temporal grouping with constraint for seam carving at t = 20

8.5. Conclusion and future works

In this chapter, a seam carving with spatio-temporal grouping constraint for video summary applications has been presented. Our proposed approach achieves a temporal rate of reduction adapted to the content, suppressing and reallocating the groups of isolated seams. An identification of important spatio-temporal groups of seams and an approximation of the number of seams by piece-wise constant segments has been introduced. By this way, the problems of salient objects geometric deformation and anachronisms have been solved. Without any constraint, the number of seams suppressed on each line is not necessary the same. With our approach, this number is defined in function of the content and the summary has the same length for all the lines. Our approach yields good results due to the fact that the constraints on the number and spatio-temporal position of suppressible are applied after having obtained the seams. This lets more flexibility to the seams and leads to a good rate of reduction, while preserving the content.

Future work will be to automatically define the spatial and the temporal thresholds on the distances, have a multi dimension detection of rupture and a correlated approximation of the number of seams by groups. The results could be also improved by using a spatio-temporal saliency maps with an object orientation. The definition of a metric to automatically evaluate the quality of the summary would also be an interesting axis of research.

9. Conclusion

In conclusion, this thesis proposes some tools to do video compression at low bitrate for security applications.

As at low bitrate, it is hard to keep all the data, it is necessary to decide what to keep and how to reduce the quantity of data in a smart way. A certain number of essentials goals depending of constraints and defined scenarios has been established. The solution has to be compatible with existing standards, define what is necessary to interpret video footage, work at low bitrate and transmit data in a flexible way, measure the influence of the suppression of the data on the semantic and work in lots of different scenarios. No constraints were done on complexity, energy consumption, camera definition or sensor type (infrared, visible, etc.). In this case, we propose a low bitrate compression method by saliency area that responds to numerous problems and conditions associated with each scenario. This will help us achieve a feasibility study.

The idea is to delete insignificant part of data contained in the video to reduce the bitrate. At first, we apply spatial resizing on the original video. Then the reduced video is encoded with a traditional encoder and data concerning spatial resizing are also encoded. At the decoder, the reduced video and data allowing to return at the original size are decoded and combined in order to return to a video with its original dimensions. The texture of the deleted areas can be synthesized if it contributes to understand the scene.

Different problematic have been seen. As it is indispensable to well detect salient objects, works on saliency maps has been done. This has leaded to two different saliency models [Décombas 2013a], [Riche 2014] and a raw video data base with manual binary mask and eyes tracking results [Riche 2014]. To reduce the quantity of information by resizing, the seam carving has been used. Different modeling has been tested giving different results presented in [Décombas 2011], [Décombas 2012a], [Décombas 2012b] and [Décombas 2014]. The different models have been presented in an international patent [Décombas 2012d].To evaluate the results of video compression by seam carving, an object based quality metric [Décombas 2012c] has been proposed and validated with subjective tests. Based on the idea that the seam carving can do spatial resizing, it has been adapted to do temporal resizing and tested to do video summary [Décombas 2013b],[Décombas 2013c].

The database proposed in [Riche 2014] has been used to evaluate the performance of the last saliency models [Riche 2014] and the last seams modeling for video compression of [Décombas 2014]. The main contributions of [Riche 2014] are a spatio-temporal rarity-based algorithm with priors for human fixations prediction and objects detection in videos called STRAP. (2a) A temporal compensation of the movement on a sliding windows allowing to manage both static and moving cameras and (2b) giving more robust spatial and temporal features is realized. (2c) These features are combined together with a new model based on rarity and low priors. (2d) High priors information are combined to the salient models to increase the performance and (2e) a segmentation model is used to have a more object based approach.

The contributions proposed in [Décombas 2014] is (6a) an algorithm that automatically cuts the sequence into GOPs, depending on the content, (6b) a spatio-temporal seam clustering method, based on spatial and temporal distances, (6c) an isolated seam discarding technique, improving the

seam encoding, (6d) a new seam modeling, avoiding geometric distortion and resulting in a better control of the seam shapes at the decoder without saliency map and (6e) a new encoder that reduces the number of bits to transmit.

The performance of the saliency model [Riche 2014] has been evaluated with different high priors information to be consistent with the other models. 7 models are chosen for the evaluation on the 3 references (eyes tracking first view, eyes tracking second view, binary mask) with 4 different metrics. In a general way, our model gives very good performances. The performance of [Décombas 2014] has been tested on Coastguard, Ducks, Parkrun, Parkjoy sequences. The Bjontegaard's metric is used to evaluate the performance of our approach. This metric allows to compute the average percentage of bitrate savings between two rate-distortion curves. For Coastguard the score is 37.27%, for Duck 19.07%, for Parkjoy 26.45% and for Parkrun 18.30%.

In the case of video summary, due to the lack of metric and pertinent database, only a few tests has been done but the preliminary results are good.

10. Perspectives

This thesis has proposed several solutions but there are still lots of challenges.

About the saliency models, it will be interesting to do the rarity on small cubes (2D+*t*), to integrate other high priors information, like cyclic phenomena. More performing tracking tools, global motion estimation could be also integrated. In continuity with the wish to have a saliency model that identify well the object, the use of object recognition will allow to have a tool that automatically identify important objects and characterize it in all the circumstances. It has also been shown that the model performs bad compare to a second viewing human prediction. Following the idea that the saliency models are usually used in more complex system, study to measure the influence of compression artifacts on saliency models could be interesting to develop a new model robust to compression artifacts.

About the video compression base on seam carving, it will be interesting to test this approach in predictive mode for H.264/AVC but also with HEVC, to improve the temporal modeling of the seams, to take into account a cost map that will highlight the areas of the sequences that need lots of bitrates, to do some experiments on the influence of letting the seams without texture synthesis or applying synthesis, to test some synthesis algorithms to reconstruct the background, to evaluate the complexity of the approach and adapting it to some products.

About the seam carving for video summary, it will be useful for the future work to automatically define the spatial and the temporal thresholds on the distances, have a multi dimension detection of rupture and a correlated approximation of the number of seams by groups. The results could be also improved by using a spatio-temporal saliency maps with an object orientation. The definition of a metric to automatically evaluate the quality of the summary would also be an interesting axis of research.

11. Bibliography

[Achanta 2009a] R. Achanta, S. Hemami, F. Estrada & S. Susstrunk, “Frequency-tuned Salient Region Detection”, IEEE Proc. on International Conference on Computer Vision and Pattern Recognition, pp. 1597 - 1604, 2009.

[Achanta 2009b] R. Achanta & S. Susstrunk, “Saliency detection for content-aware image resizing Image Processing”, IEEE Proc. on International Conference on Image Processing, pp. 1005-1008, 2009.

[Anh 2009] N. Anh, W. Yang & J. Cai, “Seam carving extension: a compression perspective”, IEEE Proc. on International Conference on Multimedia, pp. 825 - 828, 2009.

[Avidan 2007] S. Avidan & A. Shamir, “Seam Carving for Content-Aware Image Resizing”, ACM Trans. on Graphics, vol. 26, no. 10, 2007.

[Azuma 2011] D. Azuma, Y. Tanaka, M. Hasegawa & S. Kato, “SSIM based image quality assessment applicable to resized images,” IEICE Tech. Rep., vol. 110, no. 368, pp. 19-24, 2011.

[Bastani 2010] V. Bastani, M. S. Helfroush, and K. Kasiri, “Image compression based on spatial redundancy removal and image inpainting,” Zhejiang University Press, co-published with Springer, pp. 91-100, 2010.

[Belardinelli 2009] A. Belardinelli, F. Pirri, & A. Carbone, “Motion saliency maps from spatio temporal filtering”, Attention in Cognitive Systems. Springer Berlin Heidelberg, pp. 112-123, 2009.

[Bertalmío 2000] M. Bertalmío, G. Sapiro, V. Caselles and C. Ballester, “image inpainting”; IEEE Proc. of SIGGRAPH, 2000.

[Bertalmío 2001] M. Bertalmío, A. Bertozzi, G. Sapiro, “Navier-Stokes, Fluid-Dynamics and Image and Video Inpainting”, IEEE Proc. on Conference on Computer Vision and Pattern Recognition, 2001.

[Bjontegarrd 2001] G. Bjontegarrd, “Calculation of average PSNR differences between RDcurves,” in *VCEG Meeting*, Austin, USA, Apr. 2001.

[Boiman 2007] O. Boiman & M. Irani, “Detecting irregularities in images and in video”, International Journal of Computer Vision, vol. 74, no. 1, pp. 17-31, 2007.

[Borji 2011] A. Borji: Evaluation measures for saliency maps: CC_and_NSS, https://sites.google.com/site/saliencyevaluation/evaluation-measures, V.1, 2011 Extract from the web on 13.09.01.

[Borji 2013] A. Borji, & L. Itti., “State-of-the-art in visual attention modeling”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185 - 207, 2013.

[Bosch 2007] M. Bosch, F. Zhu & E. Delp, “Spatial texture models for video compression”, IEEE Proc. on International Conference on Image Processing, vol. 1, pp. 93 - 96., 2007.

[Bosch 2011] M. Bosch, F. Zhu & E. Delp, “Segmentation Based Video Compression Using Texture and Motion Models”, IEEE Proc. on Selected Topics in Signal Processing, vol. 5, no. 7, pp. 1366 - 1377, 2011.

[Bruce 2005] N. Bruce & J. Tsotsos, “Saliency based on information maximization”, In: Advances in neural information processing systems, pp. 155 - 162, 2005.

[Bruce 2009] N.D.B. Bruce & J.K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach”, Journal of Vision, vol. 9, no. 3, 2009.

[Butko 2008] N.J. Butko, L. Zhang, G. Cottrell & J. Movellan, “Visual saliency model for robot cameras”, IEEE Proc. on International Conference on Robotics and Automation, pp. 2398 - 2403, 2008.

[Carte_Reseau] http://fr.wikipedia.org/wiki/Carte_r%C3%A9seau Extract from the web on 13.09.01.

[Chambolle 2010] A. Chambolle & T. Pock, “A first-order primal-dual algorithm for convex problems with application to imaging,” Technical Report, 2010.

[Chao 2011] W.L. Chao, H.H. Su, S.Y. Chien, W. Hsu, & J.J. Ding, “Coarse-to-fine temporal optimization for video retargeting based on seam carving”, IEEE Proc. on International Conference on Multimedia and Expo, pp. 1 - 6, 2011.

[Chen 2008] B. Chen, P. Sen, “Video carving”, EUROGRAPHICS, 2008.

[Chen.Z 2010] Z. Chen, W. Lin, & K. N. Ngan, “Perceptual video coding: challenges and approaches”, IEEE Proc. on International Conference on Image Processing, pp. 784 - 789, 2010.

[Chen.H 2010] H. Chen, R. Hu, D. Mao, R. Thong, & Z. Wang, “Video coding using dynamic texture synthesis,” IEEE Proc. on International Conference on Image Processing, pp. 203 - 208, 2010.

[Chen.X 2012] X. Chen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” IEEE Proc. on Conference on Computer Vision and Pattern Recognition, pp. 853-860, 2012.

[Culibrk 2010] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, D. Kukolj, “Salient motion features for video quality assessment,” IEEE Trans. on Image Processing, 2010.

[Dailymail 2011] http://www.dailymail.co.uk/news/article-1362493/One-CCTV-camera-32-people-Big-Brother-Britain.html Extract from the web on 13.09.01.

[Décombas 2011] M. Décombas, F. Capman, E. Renan, F. Dufaux & B. Pesquet-Popescu, “Seam carving for semantic video coding ,” SPIE Proc. Application of Digital Image Processing, 2011.

[Décombas 2012a] M.Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu & F. Capman, “Improved seam carving for semantic video coding,” IEEE Proc. on MultiMedia Signal Processing, pp. 53 - 58, 2012.

[Décombas 2012b] M.Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu & F. Capman, “Closed loop seams approximation for video compression”, IEEE. Proc. On Int. symposium on Signal, Image, Video and Communications, 2012.

[Décombas 2012c] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu& F. Capman, “A new object based quality metric based on SIFT and SSIM,” IEEE Proc. On International Conference on Image Processing, pp. 1493-1496, 2012.

[Décombas 2012d] M. Décombas, E. Renan, F. Capman, F. Dufaux, B. Pesquet-Popescu, “Method for encoding an image after resizing by deleting pixels”, WO 2013014290 A1, 2012.

[Décombas 2013a] M. Décombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, & T. Dutoit, “Spatio-temporal saliency based on rare model”, IEEE Proc. on International Conference on Image Processing, 2013.

[Décombas 2013b] M. Décombas, F. Dufaux & B. Pesquet-Popescu, “Résumé vidéo par regroupement de seams”, GRETSI, 2013.

[Décombas 2013c] M. Décombas, F. Dufaux & B. Pesquet-Popescu, “Spatio-temporal grouping with constraint for seam carving in video summary application”, IEEE Proc. on International Conference on Digital Signal Processing, 2013.

[Décombas 2014] M Décombas, Y. Fellah,′ F Dufaux, B. Pesquet-Popescu, F. Capman & E. Renan,′ Seam carving modeling for semantic video coding in security applications , Submitted in IEEE Transactions on Circuits and Systems for Video Technology , 2014.

[Dempster 1977] A.P. Dempster, N.M. Laird and D. Rubin, “*Maximum Likelihood from Incomplete Data via the EM Algorithm*”, Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pp. 1 - 38, 1977.

[Deng 2011] C. Deng, W. Lin & J. Cai “Content-Based Image Compression for Arbitrary-Resolution Display Devices”, IEEE Proc. on International Conference on Communications, pp. 1 - 5, 2011.

[Domingues 2010] D. Domingues, A. Alahi & P. Vandergheynst, “Stream carving: An adaptive seam carving algorithm”, IEEE Proc. on International Conference on Image Processing, pp. 901-904, 2010.

[Dufaux 2000] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Transactions on Image Processing*, vol.9, no.3, pp.497,501, 2000.

[Facelab] Seeing Machines: Facelab commercial eye tracking system, http://www.seeingmachines.com/product/facelab/ Extract from the web on 13.09.01.

[Field 1987] D. J. Field, “Relations between the statistics of natural images and the response properties of cortical cells,” Journal of the Optical Society of America A, vol. 4, pp. 2379-2394, 1987.

[Figaro 2010] http://www.lefigaro.fr/actualite-france/2010/12/20/01016-20101220ARTFIG00493-paris-lance-son-plan-de-videosurveillance.php

[Frankovich 2011] M. Frankovich, & A. Wong, “Enhanced Seam Carving via Integration of Energy Gradient Functionals”, IEEE Trans. On Signal Processing Letters, 2011.

[Frintrop 2006] S. Frintrop, “Vocus: A visual attention system for object detection and goal-directed search”, vol. 3899, Springer, 2006.

[Gao 2008] D. Gao, V. Mahadevan & N. Vasconcelos, “On the plausibility of the discriminant center-surround hypothesis for visual saliency”, Journal of Vision, vol. 8, no. 7, pp. 1 - 29, 2008.

[Goferman 2012] S. Goferman, L. Zelnik-Manor & A. Tal, “Context-aware saliency detection”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34, no 10, pp. 1915 - 1926, 2012.

[Guo 2010] C. Guo & L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression”, IEEE Trans. on Image Processing, vol. 19, no. 1, pp. 185 - 198, 2010.

[Gupta 2011] R. Gupta & S. Chaudhury, “A scheme for attentional video compression”, Springer Pattern Recognition and Machine Intelligence, pp. 458 - 468, 2011.

[Hadizadeh 2012] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," IEEE Trans. on Image Processing, vol. 21, no. 2, pp. 898 - 903, 2012.

[Han 2012] G.J. Han, J.R. Ohm, W.J. Han & T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard”, 2012.

[Harel 2006] J. Harel, C. Koch & P. Perona, “Graph-based visual saliency”, Advances in Neural Information Processing Systems, pp. 545 - 552, 2006.

[Hou 2007] X. Hou & L. Zhang, “Saliency detection: A spectral residual approach”, IEEE Proc. on Conference on Computer Vision and Pattern Recognition, pp. 1 - 8, 2007.

[Huffman 1952] D.A. Huffman, “A Method for the Construction of Minimum-Redundancy Codes”, Proceedings of the I.R.E, pp. 1098-1102, 1952.

[Hwang 2008] D.S. Hwang & S.Y. Chien, “Content-aware image resizing using perceptual seam carving with human attention model”, IEEE Proc. on International Conference on Multimedia and Expo, pp. 1029 - 1032, 2008.

[Iffremont] http://www.ifremmont.com Extract from the web on 13.09.01.

[Irani 1996] M. Irani, P. Anandan, J. Bergen, R.Kumar, and S. Hsu, “Efficient representations of video sequences and their applications,” Image Communication Signal Processing, vol. 8, no. 4, pp. 327-351, 1996.

[Itti 1998] L. Itti, C. Koch, & E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254 - 1259, 1998.

[Itti 1999] L. Itti, C. Koch, Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems, In: Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA, Vol. 3644, pp. 473 - 482, Bellingham, WA:SPIE Press, 1999.

[Itti 2004] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention”, IEEE Trans. on Image Processing, vol. 13, no. 10, pp. 1304 - 1318, 2004.

[Itti 2006] L. Itti & P.F. Baldi, Modeling what attracts human gaze over dynamic natural scenes, in L. Harris & M. Jenkin (eds), Computational Vision in Neural and Machine Systems, Cambridge University Press, Cambridge, MA, 2006.

[ITU-R 2009] Recommendation ITU-R BT.500-12, Methodology for the subjective of the quality of television pictures, 2009.

[Iwatchlife 2013] http://www.iwatchlife.com/ Extract from the web on 13.09.01.

[Jouneau 2011] E. Jouneau & C. Carincotte, “Particle-based tracking model for automatic anomaly detection”, IEEE Proc. on International Conference on Image Processing, pp. 513 - 516, 2011.

[Judd 2011] http://people.csail.mit.edu/tjudd/TJuddThesisDefenseSlides.pdf Extract from the web on 13.09.01.

[JPEG2000] http://www.jpeg.org/jpeg2000/CDs15444.html Extract from the web on 13.09.01.

[JVT 2003] Joint Video Team of ITU-T and ISO/IEC JTC 1, “Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 ISO/IEC 14496-10 AVC),” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, 2003.

[Kang 2006] H.W. Kang, Y. Matsuhita, X. Tang, & X.Q. Chen, “Space-time video montage,” IEEE Proc. on Conference on Computer Vision Pattern Recognition, vol.2, pp. 1331 - 1338, 2006.

[Koch 1985] C. Koch, &S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” Matters of Intelligence. Springer Netherlands, pp. 115 - 141, 1987.

[Lau] B. Lau, B.: Evaluation measures for saliency maps: AUROC, http://www.subcortex.net/research/code/area_under_roc_curve Extract from the web on 13.09.01.

[Landa 2006] http://obligement.free.fr/articles/supporttransmission.php Extract from the web on 13.09.01.

[Le Meur 2006] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau, “A coherent computational approach to model bottom-up visual attention” , IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 802 - 817, 2006.

[Le Meur 2007] O. Le Meur, P. Le Callet & D. Barba, “Predicting visual fixations on video based on low-level visual features”, Vision Research, vol. 47, no. 19, pp. 2483 - 2498, 2007.

[Lepoint 2013] http://www.lepoint.fr/auto-addict/securite/videosurveillance-a-paris-pour-le-stationnement-aussi-20-03-2013-1643353_657.php Extract from the web on 13.09.01.

[Li 2009] Z. Li, P. Ishwar & J. Konrad, “Video Condensation by Ribbon Carving”, IEEE Trans. on Image Processing, vol. 18, pp. 2572 - 2583, 2009.

[Li 2011] Z. Li, S. Qin, & L. Itti, “Visual attention guided bit allocation in video compression”, Image and Vision Computing, vol. 29, no. 1, pp. 1 - 14, 2011.

[Liu 2004] Y.J. Liu, X. Luo, Y.M. Xuan, W.F. Chen, X.L. Fu, “Image Retargeting Quality Assessment,” Eurographics, vol. 30, no. 2, 2011.

[Lowe 2004] D. Lowe, “Distinctive image features from scale invariant key points,” *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91 - 110, 2004.

[Mancas] M. Mancas, N. Riche, Computational attention website http://tcts.fpms.ac.be/attention Extract from the web on 13.09.01.

[Mancas 2007a] M. Mancas, “Computational Attention Towards Attentive Computers”, Presses universitaires de Louvain, 2007.

[Mancas 2007b] M. Mancas, B. Gosselin & B. Macq, “Perceptual image representation”, Journal of Image Video Processing, vol. 2007, no. 2, pp. 3 - 3, 2007.

[Mancas 2009] M. Mancas, “Relative influence of bottom-up and top-down attention”, Attention in Cognitive Systems, Springer, vol. 5395, pp. 212 - 226, 2009.

[Mancas 2011] M. Mancas, N. Riche, J. Leroy & B. Gosselin, “Abnormal motion selection in crowds using bottom-up saliency”, IEEE International Conference on Image Processing, pp. 229 - 232, 2011.

[Mancas 2012] M. Mancas, D. De Beul, N. Riche & X. Siebert, “Human Attention Modelization and Data Reduction”, Journal of Video Compression, 2012.

[Mantratec] http://www.mantratec.com/CCTV.html Extract from the web on 13.09.01.

[Marat 2012] Sophie Marat: Ph.D. Dissertation “Visual saliency models by fusion of luminance, motion and face information for eye movements' prediction during video viewing”, 2010.

[MPEG2 1994] ITU-T and ISO/IEC JTC 1, “Generic coding of moving pictures and associated audio information – Part 2: Video,” ITU-T Recommendation H.262 – ISO/IEC 13818-2 (MPEG-2), 1994.

[MPEG4 2001] ISO/IEC JTC1, “Coding of audio-visual objects – Part 2: Visual,” ISO/IEC 14496-2 (MPEG-4 visual version 1), April 1999; Amendment 1 (version 2), February, 2000; Amendment 4 (streaming profile), 2001.

[Nam 1999] J. Nam, and A. Tewfik, “Video abstract of video,” in IEEE Proc. on MultiMedia Signal Processing, pp. 117 - 122, 1999.

[Navalpakkam 2005] V. Navalpakkam & L. Itti, “Modeling the influence of task on attention”, Vision Research, vol. 45, no. 2, pp. 205 - 231, 2005.

[Ndjiki-Nya 2012] P. Ndjiki-Nya, D. Doshkov, H. Kaprykowsky, F. Zhang, D. Bull & T. Wiegand, “Perception-oriented video coding based on image analysis and completion: A review” Signal Processing: Image Communication, 2012.

[Oh 2004] J. Oh, Q. Wen, J. Lee, and S. Hwang, “Video abstraction,” in Video Data Management and Information Retrieval, S. Deb, Ed. Hershey, PA: Idea Group, Inc./ IRM Press, pp. 321 - 346, chap. 3, 2004.

[Oliva 2003] A. Oliva, A. Torralba, M. Castelhano & J. Henderson, “Top-down control of visual attention in object detection”, IEEE Proc. on International Conference on Image Processing, vol. 1, pp. I - 253, 2003.

[Parisien 2012] http://www.leparisien.fr/paris-75/8200-cameras-sur-le-reseau-ratp-20-03-2012-1914704.php Extract from the web on 13.09.01.

[Petrovic 2005] N. Petrovic, N. Jojic & T. Huang, “Adaptive video fast forward,” Multimedia Tools Appl., vol. 26, no. 3, pp. 327 - 344, 2005.

[Popularmechanics 2009] http://www.popularmechanics.com/technology/military/4236865-3 Extract from the web on 13.09.01.

[Pritch 2008] Y. Pritch, A. Rav-Acha & S. Peleg, “Non-chronological video synopsis and indexing,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, 2008.

[Pritch 2009] Y. Pritch, S. Ratovitch, A. Hendel & S. Peleg, “Clustered Synopsis of Surveillance Video”, 6th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, 2009.

[Privitera 2000] C. M. Privitera, & L. W. Stark, “Algorithms for defining visual regions-of-interest: Comparison with eye fixations”, IEEE Trans. on Pattern Analysis and Machine Intelligence ,vol. 22, no. 9, pp. 970 - 982, 2000.

[Rahtu 2009] E. Rahtu & J. Heikkilä, “A simple and efficient saliency detector for background subtraction”, IEEE Proc. on International Workshop on Visual Surveillance, pp. 1137 - 1144, 2009.

[Ran 1995a] X. Ran, & N. Farvardin, “A Perceptually motivated three-component image model-Part I: description of the model”, IEEE Trans. on Image Processing, vol. 4, pp. 401 - 415, 1995.

[Ran 1995b] X. Ran & N. Farvardin, “A perceptually motivated three-component image model-part II: applications to image compression”, IEEE Trans. on Image Processing, vol. 4, pp. 430 - 447, 1995.

[RATP] http://www.ratp.fr/fr/ratp/c_10556/le-metro-cest-paris/ Extract from the web on 13.09.01.

[Rav-Acha 2006] A. Rav-Acha, Y. Pritch & S. Peleg, “Making a long video short: Dynamic video synopsis,” in Proc. IEEE Conf. Computer Vision Pattern Recognition, vol.1, pp. 435 - 441, 2006.

[Rav-Acha 2007] A. Rav-Acha, Y. Pritch, D. Lischinski & S. Peleg, “Dynamosaicing: Mosaicing of dynamic scenes,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1789 - 1801, 2007.

[Reeve 2011] http://www.securitynewsdesk.com/2011/03/01/how-many-cctv-cameras-in-the-uk/ Extract from the web on 13.09.01.

[Riche 2012a] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin & T. Dutoit, “Dynamic saliency models and human vision: a comparative study on videos,” IEEE Proc. on Asian Conference on Computer Vision, 2012.

[Riche 2012b] N. Riche, M. Mancas, B. Gosselin & T. Dutoit, “RARE: A new bottom up saliency model”, IEEE Proc. on International Conference on Image Processing, 2012.

[Riche 2013] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, T. Dutoit, “RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis”, Signal Processing: Image Communication, 2013.

[Riche 2014] N. Riche, M. Décombas, M. Mancas, F. Dufaux, B. Pesquet-Popescu, B Gosselin, T. Dutoit & Y. Fellah , “STRAP: A spatio-temporal rarity-based algorithm with priors for human fixations prediction and objects detection in videos”, submitted in Image and Vision Computing, 2014.

[Rijsbergen 1979] V. Rijsbergen, C. Joost "Keith", *Information Retrieval*, London, GB; Boston, MA: Butterworth, 2nd Edition,1979.

[Rubinstein 2008] M. Rubinstein, A. Shamir & S. Avidan, “Improved seam carving for video retargeting”, ACM Trans. on Graphics, Vol. 27, 2008.

[Rubinstein 2010] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, “A Comparative Study of Image Retargeting,” ACM Trans. on Graphics, vol. 29, no. 5, pp. 1 - 10, 2010.

[Sayood 2000] K. Sayood , Introduction to Data Compression, Morgan Kaufmann,Chapter 4, Section 4.4.3, 2000.

[Securiteentreprise] http://www.securiteentreprise.com/securite-temporaire.html Extract from the web on 13.09.01.

[Seo 2009] H.J. Seo, & P. Milanfar, “Static and space-time visual saliency detection by self-resemblance”, Journal of Vision, vol.9, no. 12, 2009.

[Srivastava 2008] A. Srivastava & K. Kishore Biswas, “Fast Content Aware Image Retargeting”, Sixth Indian Conference on Computer Vision, pp. 505-511, 2008

[Stentiford 2001] F.W.M. Stentiford, “An estimator for visual attention through competitive novelty with application to image compression”, Proc. on Picture Coding Symposium, pp. 25 - 27, 2001.

[Tanaka 2010a] Y. Tanaka, M. Hasegawa & S. Kato, “Image coding using concentration and dilution based on seam carving with hierarchical search”, IEEE Proc. on International Conference on Acoustics Speech and Signal Processing, pp. 1322 - 1325, 2010.

[Tanaka 2010b] Y. Tanaka, M. Hasegawa. & S. Kato, “Improved image concentration for artifact-free image dilution and its application to image coding”, IEEE Proc. on International Conference on Image Processing, pp. 1225 - 1228, 2010.

[Tanaka 2010c] Y. Tanaka, M. Hasegawa & S. Kato, “Image concentration and dilution for video coding”, Image Electronics and Visual Computing Workshop, 2010.

[Tanaka 2011a] Y. Tanaka, M. Hasegawa & S. Kato, “Seam carving with rate-dependent seam path information”, IEEE Proc. on International Conference on Acoustics Speech and Signal Processing, 2011.

[Tanaka 2011b] Y. Tanaka, M. Hasegawa & S. Kato, “Generalized selective data pruning for video sequence”, IEEE Proc. on International Conference on Image Processing, 2011.

[Tkalcic 2003] M. Tkalcic & J.F. Tasic, “Colour spaces: perceptual, historical and applicational background”, *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, vol.1, pp 304-308, 2003.

[Toet 2011] A. Toet; “Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 33, no. 11, pp.2131-2146, Nov. 2011

[Tsapatsoulis 2007] N. Tsapatsoulis, K. Rapantzikos & C. Pattichis, “An embedded saliency map estimator scheme: Application to video encoding”, International Journal of Neural Systems , vol.17, no. 4, pp. 1-16, 2007.

[Valenti 2009] R. Valenti, N. Sebe & T. Gevers, “Image saliency by isocentric curvedness and color”, IEEE Proc. on International Conference on Computer Vision, 2009.

[VGA] http://en.wikipedia.org/wiki/Video_Graphics_Array#cite_note-1 Extract from the web on 13.09.01.

[Vlahos 2008] J. Vlahos, “Welcome to the panopticon,” Popular Mechanics., vol. 1, no. 1, pp. 64–69, 2008.

[Vo 2010] D. Võ, J. Sole, P. Yin, C. Gomila& T. Nguyen, “Selective Data Pruning-Based Compression Using High-Order Edge-Directed Interpolation”, IEEE Trans. on Image Processing, vol. 19, pp. 349 - 409, 2010

[Wang.Z 2004] Z. Wang, A.C. Bovik, H.R. Sheikh & E.P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600 - 612, 2004.

[Wang.Z 2005] Z. Wang and E.P. Simoncelli, “Translation Insensitive Image Similarity in Complex Wavelet Domain” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 573 - 576, 2005.

[Wang.T 2010] T. Wang & K. Urahama, “Cartesian resizing of image and video for data compression”, IEEE Region 10 Conference, pp. 1651 - 1656, 2010.

[Welch 1984] T.A. Welch, “A technique for high-performance data compression”, Computer, vol. 17, pp. 8 - 19, 1984.

[Witten 1987] I.H. Witten, R.M. Neal, M. Radford & J.G. Cleary, “Arithmetic coding for data compression. Communications of the ACM”, vol. 30, no. 6, pp. 520 - 540, 1987.

[Wu 2005] H. R. Wu & K. R. Rao, “Digital video image quality and perceptual coding (signal processing and communications),” CRC Press, 2005.

[Xiph.org] http://media.xiph.org/video/derf/ Extract from the web on 13.09.01.

[Yeung 1997] M. Yeung, & B.L.Yeo, “Video visualization for compact presentation and fast browsing of pictorial content,” IEEE Trans. on Circuits System Video Technology, vol. 7, no. 5, pp. 771 - 785, 1997.

[Zhang 2013] L. Zhang, Z. Gu & H. Li, “SDSP: A novel saliency detection method by combining simple priors”, IEEE Proc. on International Conference on Image Processing, 2013.

[Zhu 2012] X. Zhu, D. Ramanan. “Face detection, pose estimation and landmark localization in the wild”, IEEE Proc. on International Conference on Computer Vision and Pattern Recognition, 2012.

Compression vidéo bas débit par analyse du contenu

RESUME :

L'objectif de cette thèse est de trouver de nouvelles méthodes de compression sémantique compatible avec un encodeur classique tel que H.264/AVC. . L'objectif principal est de maintenir la sémantique et non pas la qualité globale. Un débit cible de 300 kb/s a été fixé pour des applications de sécurité et de défense Pour cela une chaine complète de compression a dû être réalisée. Une étude et des contributions sur les modèles de saillance spatio-temporel ont été réalisées avec pour objectif d'extraire l'information pertinente. Pour réduire le débit, une méthode de redimensionnement dénommée «seam carving » a été combinée à un encodeur H.264/AVC. En outre, une métrique combinant les points SIFT et le SSIM a été réalisée afin de mesurer la qualité des objets sans être perturbée par les zones de moindre contenant la majorité des artefacts. Une base de données pouvant être utilisée pour des modèles de saillance mais aussi pour de la compression est proposée avec des masques binaires. Les différentes approches ont été validées par divers tests. Une extension de ces travaux pour des applications de résumé vidéo est proposée.

Mots clés : Seam carving, H.264 / AVC, saliency, metric, SIFT, video summary, manual binary mask

Low bitrate video compression by content characterization

ABSTRACT :

The objective of this thesis is to find new methods for semantic video compatible with a traditional encoder like H.264/AVC. The main objective is to maintain the semantic and not the global quality. A target bitrate of 300 Kb/s has been fixed for defense and security applications. To do that, a complete chain of compression has been proposed. A study and new contributions on a spatio-temporal saliency model have been done to extract the important information in the scene. To reduce the bitrate, a resizing method named seam carving has been combined with the H.264/AVC encoder. Also, a metric combining SIFT points and SSIM has been created to measure the quality of objects without being disturbed by less important areas containing mostly artifacts. A database that can be used for testing the saliency model but also for video compression has been proposed, containing sequences with their manually extracted binary masks. All the different approaches have been thoroughly validated by different tests. An extension of this work on video summary application has also been proposed.

Keywords : Seam carving, H.264 / AVC, saliency, metric, SIFT, video summary, manual binary mask

TELECOM ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY