# Annotation de comportements à risques et suivi de visages dans les remontées mécaniques en montagne.

Cyril Meurie, Rémi Dufour, Amaury Flancquart

Univ Lille Nord de France, F-59000 Lille IFSTTAR, COSYS, LEOST, F-59650 Villeneuve d'Ascq
cyril.meurie@ifsttar.fr, remi.dufour@ifsttar.fr, amaury.flancquart@ifsttar.fr

Résumé – Cet article vise à promouvoir le projet EVEREST (Évaluation des pERformances des systèmes vidÉo pour la Sécurité des Transports guidés en montagne) dont la finalité consiste, à évaluer, lors d'un challenge qui sera organisé en 2020, les performances des systèmes à base d'analyse d'images proposés par des participants, à détecter des comportements potentiellement dangereux d'usagers de transports guidés en montagne. Les outils d'annotation et de détection notamment de visages disponibles dans la littérature s'avèrent inadaptés pour notre application. Nous proposons donc un outil intuitif et flexible d'annotation semi-automatique intégrant une stratégie de suivi automatique de visages (combinant un CAMShift et un flux optique) afin de faciliter le travail d'annotation des usagers et de leurs visages sur les 96 heures (10.368.000 images) de données vidéos réelles acquises sur trois sites expérimentaux.

Abstract – This paper aims, firstly, to promote the EVEREST project, which aims to evaluate, during a challenge which will take place in 2020, a video based systems proposed by competitors, consisting of detecting hazardous behaviors of users of ski lifts. This research domain is very specialized and has not been explored deeply yet, which is why existing annotation software and detection algorithms are not suitable. Secondly, we propose a new intuitive semi-automatic annotation software, as well as a new face tracking method which is robust to the specific challenges of the EVEREST project.

## 1 Introduction

Depuis plusieurs années, les autorités s'intéressent aux performances de certains dispositifs de sécurité dédiés à la surveillance de l'embarquement et du débarquement des transports guidés en montage. Le rapport Ligeron (2013) montre que le nombre d'accidents est plus élevé à l'embarquement et au débarquement et que certains types de transports guidés sont plus sujets aux accidents. Sur les systèmes actuels, la sécurité est généralement garantie par l'utilisation de capteurs binaires et la présence d'un surveillant. Toutefois, l'analyse des rapports annuels d'accidents conclut que la majorité des accidents sont causés par un comportement dangereux ou inapproprié des usagers. Face à l'évolution des technologies de vision par ordinateur, des systèmes de vidéo-surveillance émergent pour assurer des fonctions de sécurité, notamment pour la surveillance de la manoeuvre du garde-corps [1]. Depuis une dizaine d'années, des projets d'évaluation sont régulièrement proposés (PETS, CAVIAR, ETISEO, iLids, BEHAVE). Ils utilisent généralement le même protocole, les même critères de performance et mettent à disposition une base de données vidéo spécifique sur laquelle les systèmes sont exécutés. Mais aucun d'entre eux ne traite le cas des remontées mécaniques. Le projet EVEREST s'inscrit dans ce contexte d'évaluation et vise à anticiper la mise en œuvre des solutions technologiques futures qui seront mises sur le marché dans les années à venir. Dans ce papier, nous présentons le concept du projet EVEREST, illustré sur la figure 1 avec le challenge qui sera proposé à la communauté en 2020 ainsi qu'un outil d'annotation semi-automatique facilitant le travail d'annotation des comportements potentiellement dangereux des usagers et l'anonymisation de ces derniers grâce à une stratégie de suivi automatique de visages. Le projet EVEREST est une activité de recherche qui vise à estimer la capacité des systèmes vidéo à assurer les fonctions d'aide à l'exploitation et de surveillance des transports guidés en montagne. Nous avons créé une base de données vidéo (BdV) de 96 heures à partir de séquences d'images acquises par des caméras avec des vues différentes sur trois sites expérimentaux. La BdV contient des comportements potentiellement dangereux générés par des usagers ou simulés par des acteurs à savoir : i) la présence d'un usager après débarquement ; ii) le mauvais positionnement d'un usager (sous-marinage) après embarquement ; iii) un garde-corps non abaissé après embarquement. Cette base de données vidéo est annotée et divisée en deux sous-bases (Test et Validation) pour : i) permettre aux parties prenantes de proposer et d'optimiser sur la BdV-Test des algorithmes de détection de comportements à risques ; ii) nous permettre d'évaluer sur la BdV-validation tenue secrète, la performance des algorithmes proposés par les candidats au challenge qui sera organisé en 2020.

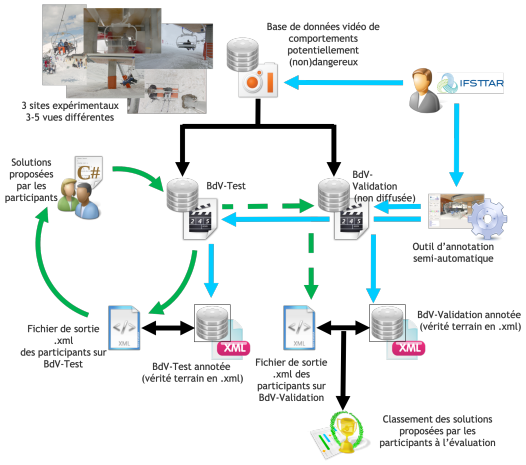


Figure 1 – Concept du projet EVEREST donnant lieu au challenge qui sera proposé à la communauté en 2020

# 2 État de l'art

Récemment, avec l'arrivée des bases de données vidéo contenant des millions d'exemples, plusieurs outils d'annotation ont été développés (VATIC, VIA [5], etc.). Ils sont essentiellement conçus pour des sujets génériques : classification, détection, ou segmentation d'objets ou de personnes en environnement urbain. Ces outils s'avèrent donc inadaptés aux besoins spécifiques du projet EVEREST. En effet, nous devons disposer d'un outil simple à prendre en main, flexible, multi-plateformes et optimisé pour une annotation et une anonymisation aisée de notre base de donnée vidéo. Nous nous sommes d'abord intéressés au défi de la détection de visages. Ce domaine de recherche a été très actif durant la dernière décennie. L'algorithme Viola-Jones [6] fut une percée importante dans ce domaine. Aujourd'hui, l'apprentissage profond a permis d'atteindre un nouveau niveau de performances. Nous avons donc évalué, sur notre base de données, l'implémentation OpenCV des algorithmes Viola-Jones et SSD [8] pour la détection de visages. Dans les deux cas, nous avons rapidement conclu que les algorithmes n'offraient pas les performances requises, A titre d'exemple, dans la plupart des cas, sur une vue de 3/4, Viola-Jones ne détecte pas tout les visages et SSD n'en détecte aucun. Cet échec s'explique par le contexte particulier de notre base de données vidéo. En effet, les usagers portent dans la plupart des cas des casques, des bonnets, et des lunettes de skis, ce qui rend la détection très difficile. Face à ce constat, et partant de l'hypothèse qu'une première détection manuelle du visage serait réalisée, nous nous nous sommes orientés vers l'utilisation d'algorithmes de suivi de visages afin de ne plus avoir à les détecter manuellement tout au long de la séquence. Nous avons alors testé une implémentation multi-objets de l'algorithme TLD (MOTLD) [4]. Les résultats étaient prometteurs mais nous avons décelé une instabilité en cas de rotation ou de rapprochement du visage suivi. Nous avons alors testé l'algorithme CAMShift qui se veut être plus robuste aux rotations mais comporte aussi des défauts. En effet, il se base sur un histogramme de la teinte de l'objet suivi, qui a ses limites quand le fond a une teinte similaire. Des efforts pour palier cette faiblesse ont déjà été explorés, tel que l'utilisation de caractéristiques de textures plutôt que de teinte [10], ou l'utilisation d'une soustraction de fond. Ces adaptations nécessitent une image de fond de référence, ce qui n'est pas envisageable dans notre contexte de remontées mécaniques, car les télésièges sont en mouvement et ne sont donc pas filtrés par la soustraction de fond.

# 3 Méthode proposée

## 3.1 Interface graphique d'annotation

Une interface graphique multi-plateformes a été développée en Qt afin de faciliter la lourde tâche d'annotation de quelques 96 heures de vidéo (10.368.000 images). Contrairement aux outils existants dans la littérature, celle-ci permet de réaliser deux tâches : l'annotation image par image, par un expert, de différents évènements et des visages des usagers. Un module optionnel de suivi de visage permet de choisir parmi trois méthodes : CAMShift [2], MOTLD [4], et la méthode proposée décrite ci-après. L'initialisation de la méthode a été réalisée manuellement, sur chaque première image d'un évènement, par la mise en place d'une boite englobante (via un clic) centrée sur le visage de l'usager. Les différents évènements annotés peuvent correspondre aux trois types de comportements potentiellement dangereux mentionnés en section I ou à des événements annexes comme la marche/arrêt de l'installation, l'occultation de la scène, etc. Un fichier de sortie au format XML est généré à l'issue de l'annotation et contient pour chaque image, la présence d'évènements ainsi que la taille et les cordonnées (x,y) des boites englobantes centrées sur le visage des usagers. Au vue de la taille de la base de données, chaque séquence est annotée par un seul expert. L'utilisation de différentes vues (majoritairement synchronisées) permet de détecter d'éventuel conflit et donc de robustifier l'annotation des évènements.

## 3.2 Stratégie de suivi de visages

L'analyse des performances de l'algorithme CAMShift sur notre base de données vidéo nous a poussé à explorer l'intégration d'une information de mouvement. Naturellement, l'utilisation du flux optique nous a semblé une approche intéressante. Les algorithmes les plus couramment utilisés sont Farnebäck [3] et Lucas-Kanade [11]. Farnebäck permet d'obtenir un flux optique dense (pour chaque pixel de l'image) alors que Lucas-Kanade permet d'obtenir un flux optique creux (pour une partie des pixels de l'image). L'algorithme Farnebäck a été choisi pour sa

simplicité d'utilisation et d'intégration.
L'idée de la stratégie proposée est de combiner le flux optique au CAMshift. L'estimation du mouvement de l'objet permet de générer un masque, qui est appliqué à l'image de distribution de probabilité servant à exécuter CAMShift. De cette manière, CAMShift ne déviera pas trop de l'objet suivi, y compris lorsque le fond est de la même teinte que l'objet suivi. Cette stratégie est illustrée sur les Figures 2 et 3. Les blocs de flux optique sont représentés en bleu, tandis que ceux du CAMShift sont en noir. Les blocs de couleur verte représentent la phase d'initialisation de l'algorithme : la première boîte englobante $T_0$ et l'histogramme de l'objet suivi. La région d'intérêt à partir de laquelle l'histogramme est calculé doit être représentative de la teinte de l'objet à suivre.
Un vecteur de mouvement est calculé par la moyenne des vecteurs de flux optique à l'intérieur de la boîte englobante $T_{-1}$, qui ont une norme supérieure à un seuil fixé expérimentalement à 2. En fonction des contraintes de vitesse de l'objet ou du « framerate », ce vecteur de mouvement peut être associé à un coefficient (fixé par défaut à 1) permettant de l'adapter à la vitesse des objets à suivre. La boite englobante $T_{-1}$ est étendue dans la direction du vecteur et un masque binaire est généré pour l'image $T$ qui prend donc en compte la direction et la vitesse de l'objet suivi. Enfin, le masque est appliqué au résultat de la rétroprojection et filtre ainsi l'essentiel des potentielles distractions du fond.

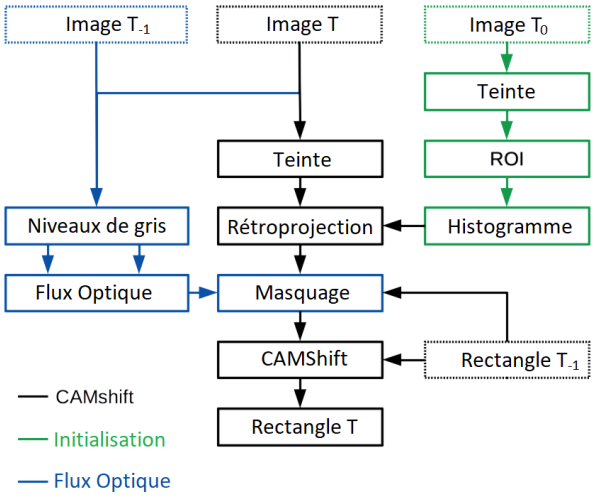


Figure 2 – Stratégie proposée combinant CAMShift et Flux optique.

## 4 Résultats expérimentaux

La performance de la méthode de suivi semi-automatique proposée (combinant CAMShift et Flux Optique) a été testée sur 3 séquences contenant respectivement 100 images pour les vues de face et de trois-quart et 57 images pour la vue de profil avec un seul usager présent sur le télésiège (majorité des cas simulés). Les résultats sont indiqués dans la Table 1. Concernant la vue de face, la méthode proposée obtient un score IoU moyen de 76% comparé respectivement à 77% pour l'algorithme MOTLD et 7% pour l'algorithme CAMShift original. Concernant la vue de trois-quart, l'approche proposée obtient le meilleur score IoU moyen, à savoir 70%, comparé à 35% pour l'algorithme MOTLD et 34% pour l'algorithme CAMShift original. Enfin, pour la vue de profil, l'approche proposée obtient également le meilleur score IoU moyen c'est-à-dire 80% comparé à 55% pour l'algorithme MOTLD et 60% pour l'algorithme CAMShift original.

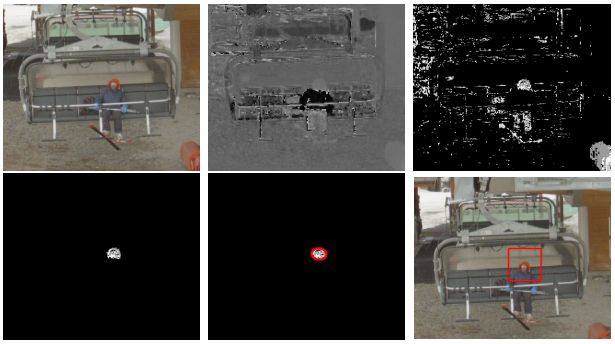

Figure 3 – Étapes de la méthode proposée (de haut en bas et de gauche à droite : $image_T$, image de teinte, rétroprojection, rétroprojection avec flux optique, ellipse obtenue par CAMShift, boite englobante résultante).

Table 1 – Performance (IoU moyen) de la méthode proposée et de deux méthodes de l'état de l'art

| IoU | Vue face | Vue 3/4 | Vue profil |
|---|---|---|---|
| CAMshift [2] | 7 % | 34% | 60% |
| MOTLD [4] | 77% | 35% | 55% |
| Notre méthode | 76 % | 70% | 80% |

Les Figures 4 et 5 illustrent de manière plus détaillée les performances mesurées. La Figure 4 détaille le score IoU moyen des trois méthodes de suivi de visages (CAMShift, MOTLD et la méthode proposée) sur l'ensemble des images des trois séquences testées. Nous pouvons remarquer qu'après initialisation, la méthode proposée suit correctement le visage des usagers sur les trois séquences considérées, alors que l'algorithme CAMShift diverge dès la $10^{me}$ image pour la vue de face et dès la $25^{me}$ image pour la vue de trois-quart. l'algorithme MOTLD obtient un résultat comparable à notre méthode pour la vue de face, mais diverge dès la $19^{me}$ image pour la vue de trois-quart et dès la $35^{me}$ image pour la vue de profil.

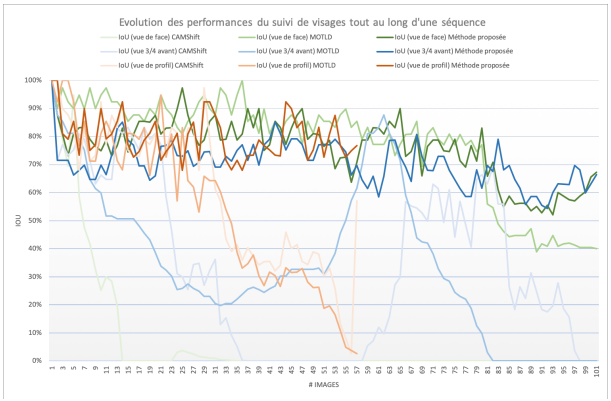


Figure 4 – Évolution des performances de CAMshift, MOTLD et de la méthode de suivi de visage proposée sur 3 séquences présentant des vues différentes.

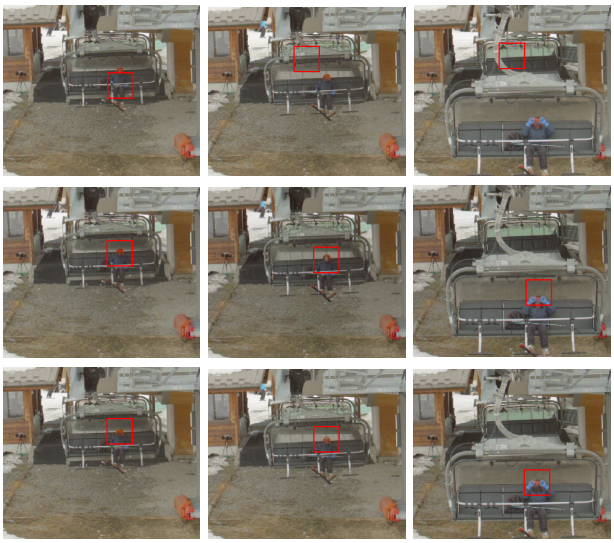

Figure 5 – Évolution des performances du suivi de visage aux images #10, #25 et #100 (de haut en bas : CAMshift, MOTLD et la méthode proposée).

# 5 Conclusion

Dans cet article, nous faisons, d'une part, la promotion du projet EVEREST qui vise à évaluer les performances des algorithmes de détection de comportements potentiellement dangereux d'usagers de transports guidés en montagne qui seront proposés par les candidats au challenge qui sera organisé en 2020 au travers d'un workshop. Nous proposons également un nouvel outil intuitif, semi-automatique d'annotation de vidéos, conçu spécialement pour les besoins du projet EVEREST. Cet outil permet de créer la vérité terrain des trois types de comportements à risques des usagers de remontées mécaniques, et inclut des fonctionnalités de suivi automatique de visages (après initialisation) en vue de leur anonymisation. Pour ce faire, nous proposons une nouvelle méthode de suivi de visage basée sur une combinaison des algorithmes CAMShift et Farnebäck (Flux Optique). Celle-ci s'avère plus précise et robuste avec un score IoU supérieur à 70%, comparable ou supérieur aux deux méthodes de l'état de l'art testées (CAMShift, MOTLD).

# Remerciements

Ce travail a été réalisé dans le cadre du projet EVEREST. Cette action de recherche fait l'objet d'un soutien financier de la part du Service Technique des Remontées Mécaniques et des Transports Guidés (STRMTG).

# Références

[1] K. Bascol, R. Emonet, E. Fromont and R. Debusschere, Improving Chairlift Security with Deep Learning, Advances in Intelligent Data Analysis XVI. LNCS 10584. Cham, pp. 1-13, 2017

[2] G-R. Bradski, Computer Vision Face Tracking For Use in a Perceptual User Interface, Intel Technology Journal, 1998.

[3] G. Farnebäck, Two-Frame Motion Estimation Based on Polynomial Expansion, SCIA. LNCS 2749. Heidelberg, pp. 363-370, 2003.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas, Tracking-Learning-Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34(7), pp. 1409-1422, 2012

[5] A. Dutta, A. Gupta and A. Zissermann, VGG Image Annotator (VIA), *http* : *//www.robots.ox.ac.uk/vgg/software/via/*, 2016

[6] P-A. Viola and M-J. Jones, Rapid object detection using a boosted cascade of simple features, CVPR, pp. 511-518, 2001.

[7] J. Redmon and A. Farhadi, YOLOv3 : An Incremental Improvement, arXiv, 2018

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A-C. Berg, SSD : Single Shot MultiBox Detector, ECCV, 2016

[9] Y. Cheng, Mean Shift, Mode Seeking, and Clustering, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 17, pp. 790-799, 1995.

[10] J. Yin, Y. Han, J. Li and A. Cao, Research on Real-Time Object Tracking by Improved Camshift, International Symposium on Computer Network and Multimedia Technology, pp. 1-4, 2009

[11] B-D. Lucas and T. Kanade, An Iterative Image Registration Technique with an Application to Stereo Vision, Darpa IU Workshop, pp. 121-130, 1981