Annotation tool designed for hazardous user behavior in guided mountain transport

Rémi Dufour, Cyril Meurie, Amaury Flancquart

Annotation tool designed for hazardous user behavior in guided mountain transport

1st Rémi Dufour
Univ Lille Nord de France, F-59000 Lille
IFSTTAR, COSYS, LEOST, F-59650 Villeneuve d'Ascq
remi.dufour@ifsttar.fr

2nd Cyril Meurie
Univ Lille Nord de France, F-59000 Lille
IFSTTAR, COSYS, LEOST, F-59650 Villeneuve d'Ascq
cyril.meurie@ifsttar.fr (Corresponding author)

3nd Amaury Flancquart
Univ Lille Nord de France, F-59000 Lille
IFSTTAR, COSYS, LEOST, F-59650 Villeneuve d'Ascq
amaury.flancquart@ifsttar.fr

***Abstract*—This paper proposes a semi-automatic ground truth annotation software designed for the specific needs of the EVEREST project. The purpose of this project is to build an annotated and anonymized video database, and use it to evaluate algorithms in the task of detecting hazardous behavior in guided mountain transport. To do so, a ground truth annotation tool that disposes designed specifically for the EVEREST project was needed. Ski lifts safety based on intelligent video systems is a niche domain which has not yet been explored in depth, which means no annotation tool suited for this task was available. That is why, we decided to develop a user-friendly and flexible tool to allows the semi-automatic annotation of events and faces (for privacy purposes). We looked at existing tracking algorithms, chose an implementation of TLD, and designed a new tracking algorithm that could be used when TLD isn't effective. This led to a simple, lightweight tracking algorithm that is more practical to use than the original CAMshift algorithm, and a user-friendly and flexible annotation tool that is well adapted to the specific task of annotating hazardous behavior in guided mountain transport.**

***Index Terms*—image processing tool, guided mountain transport, security, face tracking, annotation software, video dataset.**

I. Introduction

In recent years, authorities have been questioning the performance of certain security devices dedicated to monitoring the embarkation and disembarkation of guided mountain transport. The Ligeron report (2013) denotes on the one hand that the number of accidents (between 2002 and 2011) is higher when boarding or disembarking and on the other hand that some types of guided transports are more accident-prone. On current systems, safety is guaranteed by the use of binary sensors sometimes associated with a safety programmable logic controller (PLC) as well as the presence of a supervisor on each ski lift. However, the analysis of the annual accident reports concludes that the majority of accidents are caused by hazardous or inappropriate user behavior. That is why, the feasibility of video surveillance systems are being studied for the purpose of providing operational support functions (guardrail manoeuvre monitoring). Faced with the evolution of camera technologies and image interpretation algorithms, intelligent video surveillance systems are starting to emerge to provide security functions [1]. Any system required to provide safety functions must comply with the applicable standards in the field of cable installations and specialized standards. For the past ten years, evaluation projects have been regularly proposed (PETS, CAVIAR, ETISEO, iLids, BEHAVE, etc). Overall, they use the same protocol and identical performance metrics and make available a specific dataset on which the tested systems are executed. But none of them concern ski lifts. The EVEREST project is part of this evaluation context and aims to anticipate the implementation of such technological solutions that will go to market in the coming years. In this paper, we present on the one hand, the EVEREST project and the challenge that will be proposed to the scientific community in 2019, and on the other hand, the semi-automatic ground truth annotation software that makes it easier to annotate the three types of potentially dangerous user behavior and anonymize passengers using the proposed face tracking approach in the entire real video database (i.e 96 hours or 10,368,000 images).

The EVEREST project (ie. the acronym for vidEo surVeillance pERformancE analySis for mountain guided Transport) is a research activity that aims to estimate the ability of video based systems to ensure operating aids and monitoring functions for guided mountain transport. Figure 1 illustrates the EVEREST project's concept. To achieve this goal, we have created a 96-hours real video database from several sequences, acquired from different points of views at the ski lift departure/arrival areas of three french ski resorts. The acquired sequences contains different potentially dangerous situations generated by hazardous user behavior or simply simulated by actors. This video database is annotated and divided into two sub-databases (respectively called BdV-Test and BdV-Validation): i) to share the BdV-Test and

The people faces in the figures of this paper were deliberately blurred out of respect for privacy, but the algorithms were tested on the raw images acquired as part of the EVEREST project.

allow stakeholders to propose and optimize video based algorithms for the detection of risk situations; ii) to allow us to evaluate on the BdV-Validation that is kept secret, the performance of the proposed algorithms by the challenger candidates.

The paper is organized as follow : Section II introduces the problematic of the ground truth annotation of large databases, the difficulties to use existing software products in the context of hazardous user behavior in guided mountain transport, and a selection of well known state of the art methods in the context of face detection and tracking. In section III, we propose a tracking strategy combining CAMshift [2] and Optical flow [3] and the user-friendly and flexible Graphical User Interface developed specially for the needs of the EVEREST project. Section IV details the EVEREST database and the experimental results obtained on three sequences with different points of view and compares the proposed approach with the original CAMshift and the more recent TLD [4].

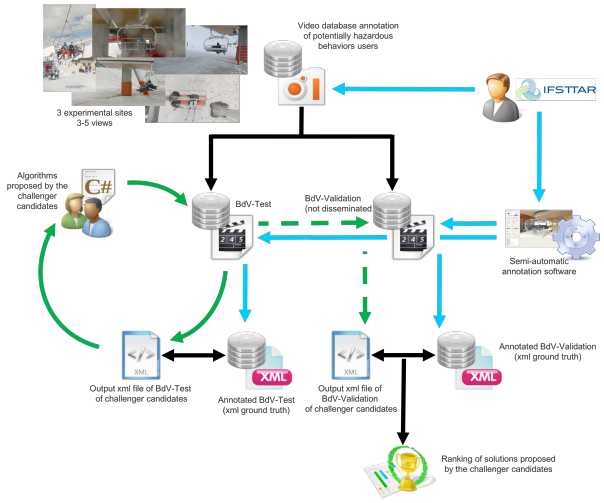

Fig. 1: EVEREST project's concept.

In this context, the vision based algorithms that will be proposed by the challenger candidates at the end of 2019 should be able to detect three types of hazardous user behavior: i) Presence of passenger beyond the ski lift arrival area (Figure 2 left); ii) Bad position of passenger from the ski lift departure area: the buttocks are not placed on the seat or the knees are at the footrest (Figure 2 middle); iii) Guardrail is not lowered completely beyond a given area with presence of passenger on the seat (Figure 2 right).

II. Literature review

A. Annotation tools

With the arrival of large image databases and the needs in terms of training and evaluation of image processing algorithms, many annotation tools have been developed in past years such as VATIC [5], VIA [6], iVAT [7], Scalabel, BeaverDam [8], CVAT. They are mainly developed for generic tasks (object detection and classification in urban

Fig. 2: Three types of hazardous user behavior (left to right: presence of passenger beyond the ski lift arrival area, bad position of passenger from the ski lift departure area, guardrail is not lowered completely beyond a given area with presence of passenger on the seat.

environments, segmentation task, etc) that are not compatible with our objective to annotate hazardous events and also anonymize people in order to respect their privacy. Indeed, VATIC is an online video annotation tool for computer vision research that crowdsources work to Amazon's Mechanical Turk and is not compatible with the privacy requirements of the EVEREST project. CVAT is not adapted to our use case and less user-friendy enough compared to our proposed tool. For example it requires to convert the full video into individual images and it is tedious for our 10,368,000 images. BeaverDam seems to be hard to use on every platform we use. In conclusion VATIC, VIA, iVAT, Scalabel, BeaverDam, CVAT seem to be not user-friendly enough and less adapted to our specific use case (multiplateforms, annotation speed and ease, privacy requirements, etc) than our proposed semi-automatic annotation tool.

B. Face detection and tracking

Face detection has been an active research subject for the last few decades. The Viola-Jones algorithm [9] was a breakthrough that offered good performances in the detection of human faces. Now, Deep Neural Networks (DNN) are the center of interests, in particular with the famous YOLO algorithm [10] and the use of Single Shot MultiBox Detector (SSD) [11]. We evaluated the OpenCV implementation of the Viola-Jones and SSD algorithms. For Viola-Jones, we tested the cascades called *haarcascade_frontalface_default* and *haarcascade_profileface*. For SSD, we used a model that was pre-trained to detect faces. Unfortunately, we found that both types of algorithms didn't provide good detection performances for our specific video database and requires for DNN approaches a good GPU that annotators do not necessarily have at the present time. Indeed, Figure 3 illustrates a random EVEREST's example where, for a front or a profile view, no face has been detected either with Viola-Jones or with SSD, and for 3/4 view, only one face out of two was detected with Viola-Jones while SDD did not detect anything. Users of ski lift often wear sunglasses, helmets and hoods, hiding most of their face, which greatly decreases the effectiveness of pre-trained detection algorithms. Since detection algorithms are unreliable for automatic anonymization, a semi-automatic face annotation workflow allowing a annotator to click on people faces in order to initialize a selected tracking algorithms.

Fig. 3: (No)Detection of people faces with Viola-Jones algorithm and SSD approaches on EVEREST database with different points of view.

We looked at the state of the art in this domain, and used a multi-object implementation of Tracking Learning Detection (MOTLD) [4]. MOTLD proved to work well in most cases, but it could be unreliable when it comes to resilience to rotations, scales and posture changes. That is why, we decided in our annotation tool to propose to the annotator to select between two tracking algorithms : MOTLD and CAMshift described below (their number can be increased in the future). The CAMshift algorithm [2] is a fast face tracking algorithm based on the Meanshift [12] algorithm, designed specifically for face tracking. Meanshift is an iterative algorithm that moves a constant size window over a probability distribution, climbing the gradient until it becomes stable. It is lacking as a tracking solution because of the constant window size, which is a handicap for tracking objects that are moving away or toward the camera. CAMshift resolves this problem, by updating at each step, not only the position of the window, but also its size and orientation. In practical use, CAMshift has been quite successful, in particular, it is quite resilient to orientation and scale changes due to its design. Nevertheless, shortcomings remain. In particular, its performance can be negatively influenced if the background of the tracked object has a similar Hue. Attempts to circumvent this weakness have been tried, such as using texture information rather than the Hue [14]. Another way to make the CAMshift algorithm more robust is to take advantage of movement data. One way which was already explored is to use background subtraction in order to restrict CAMshift to the moving parts of the image. This technique works well, but requires a reference background image, and it is not adapted to the task of tracking a user riding a chairlift, because the chairlift is an object that is not part of the background, which means it would be considered as part of the tracked object.

III. Proposed method

A. Proposed tracking approach

As indicated above, CAMshift does not offer satisfactory results on our application. This is why another way to make CAMshift-based object tracking more robust using movement data is needed. The solution explored in this paper is using optical flow. Optical flow is a field of motion vectors of every points, from the perspective of the observer [13]. The most common ways to compute it today are the Farnebäck algorithm [3], which produces a dense optical flow (for every pixel in the image), and the Lucas-Kanade Algorithm [15], which produces a sparse optical flow. The basic idea of the method proposed in this paper is to use an optical flow algorithm to estimate the movement of the object to mask the probability distribution image, so that the region surrounded by CAMshift does not deviate too much from the object in case the background is of a similar Hue. In this paper the Farnebäck algorithm was used, for its simplicity of use and integration thanks to its dense characteristic. The method is illustrated in Figure 4. The general case of the method is marked in blue (for the optical flow block) and black (for the CAMshift block), while the green part is done only in the initialization step (ie. T0). Two parts need to be initialized, the first bounding box T0, and the histogram that will be used to perform backprojection. The Region of Interest (ROI) from which the histogram is computed is meant to represent the Hue to be tracked. As illustrated in Figure 5 (left) and for our application, the annotator can select a part of the face to track as the ROI. The proposed approach is illustrated in the synopsis of the Figure 4. The core part of the method, using the optical flow to help guide CAMshift, is the Mask step (as seen in Figure 4). It consists of generating a mask using the previous frame's bounding box and the output of the optical flow, and applying it to the backprojection. The mask is generated thanks to a displacement vector that extends the bounding box's boundaries as shown in Figure 5 (right). This displacement vector is computed by averaging the optical flow vectors that are inside the previous bounding box and that are higher than a threshold, which leads to a vector that roughly approximate the motion of the tracked object. Depending on the need to track high speed object or an eventual low framerate constraint, the coefficient w governing the influence of the motion vector on the shape of the extended bounding box can be increased from its default value of 1. The extended bounding box is used to generate a binary mask that is applied to the backprojection.

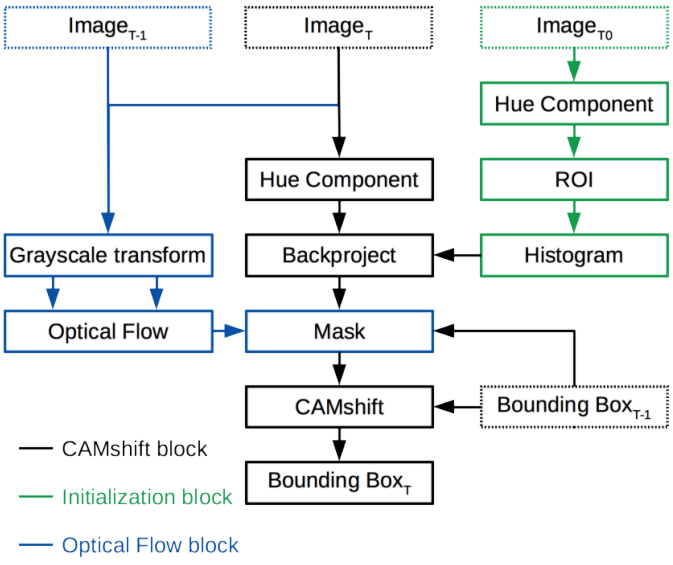

Fig. 4: The proposed method combining CAMshift and Optical flow.

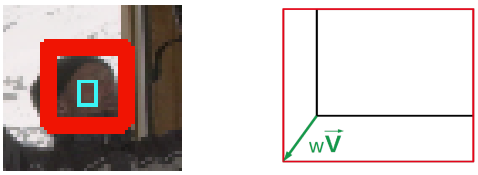

Fig. 5: Illustration of a bounding box (marked in red) and a ROI (marked in blue) defined manually by an annotator (left), Bounding box extension by the average displacement vector (right)

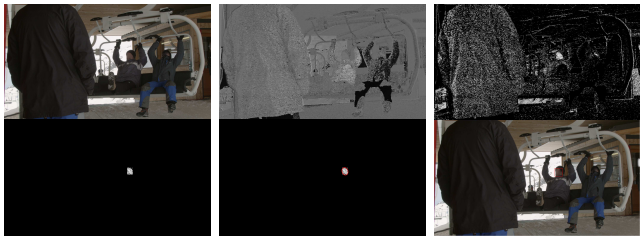

Fig. 6: Different steps of the proposed method (from left to right and top to bottom: $frame_T$, RGB to Hue transformation, backprojection, backprojection after applied mask, ellipse obtained by CAMshift, resulting bounding).

B. The Graphical User Interface

A Graphical User Interface (GUI) illustrated in Figure 7 was developed in order to facilitate the very time-consuming annotation of 96 hours of video (ie. 10,368,000 images). It specializes in two types of annotations: events and bounding boxes. Events are meant to cover all the three hazardous events mentioned in section II, plus additional ones, meant for producing negative events examples, indicating the complete stop of the guided mountain transport, or an occlusion of the camera. Bounding boxes are meant to be placed on potentially recognizable faces in the Region of Interest (ROI) of the event being annotated. Their size has been specified to be constant for each particular video in the database. Placing bounding boxes manually for the whole database would have been a very time consuming project, which is why a tracking algorithm adapted to the particular case of the EVEREST video database was developed.

In order to disseminate the video database, it is required to anonymize the users of guided mountain transport. To do so, bounding boxes around their faces are needed so that they can be blurred. However, most face detection algorithms have been tested and do not offer satisfactory results due to the specific conditions (ski goggles, ski helmet, cap, hoods, etc). That is why, we have chosen to initialize manually the position of user face and track it over several frames with an optimized algorithm.

IV. Experiments

A. Database and evaluation procedure

The real video database that we have created contains 96 hours with a framerate of 30fps. This database is divided into two subdatabases. We recall that the BdV-Test will be available for the challenger candidates and the BdV-Validation kept secret and used to evaluate the candidate algorithms. The size of the acquired image is 1920×1080 and sequences are acquired by several positions of cameras (front view, 3/4 view, profile view, back view and top view).

Fig. 7: The annotation software designed.

In order to evaluate the performance of the tracking algorithms, ground truth annotations were created manually. Then, a comparison to the ground truth was done based on two metrics: Intersection over Union (IoU) and center distance. IoU is a simple and common metric for evaluating how well two bounding boxes match. It simply consists of dividing the intersection of the two bounding boxes by their union. It produces a value between 0 and 1 (also converted in percentage) that reflects how well the two bounding boxes match. Its advantage lies in discriminating well between bounding boxes that are of different sizes. Sometimes, we care more about the position of the center of the bounding box than its size and this evaluation method can be misleading in that regard. This is why, a second metric, the center distance between two bounding boxes, was used to compensate for this weakness of IoU. Our evaluation procedure takes the average values of both these metrics when comparing the results of the algorithm with a ground truth reference database.

B. Experimental results

The performance of the proposed approach (CAMshift coupled with Optical Flow) has been tested on three sequences containing 100 frames (ie. approximately 3 seconds) for front view and 3/4 view and 57 frames (ie. approximately 1 second) for profile view. As indicated in Tables I and II, in the first sequence (ie. front view), the proposed approach obtained a similar average IoU score of 76% compared to 77% obtained with the MOTLD algorithm but better than the 7% obtained

with the CAMshift used alone. In the second sequence (ie. 3/4 view), the proposed approach obtained the best average IoU score of 70% compared to 35% obtained with the MOTLD algorithm or 34% obtained with the CAMshift used alone. In the third sequence (ie. profile view), the proposed approach also obtained the best average IoU score of 80% compared to 55% obtained with the MOTLD algorithm or 60% obtained with the CAMshift used alone. In addition and for all the sequences, the center distance obtained with the proposed approach is similar or smaller than the distance obtained with the MOTLD algorithm or the CAMshift used alone.

TABLE I: Average IoU performance of the proposed method compared with two state-of-the-art methods

IoU	**Front view**	**3/4 view**	**Profile view**
CAMshift [2]	7 %	34%	60%
MOTLD [4]	**77%**	35%	55%
Our method (CAMshift + OF)	76 %	**70%**	**80%**

TABLE II: Average center distance performance of the proposed method compared with two state-of-the-art methods

IoU	**Front view**	**3/4 view**	**Profile view**
CAMshift [2]	140 px	58 px	19 px
MOTLD [4]	**9 px**	48 px	23 px
Our method (CAMshift + OF)	10 px	**11 px**	**7 px**

A more complete performance overview is given in Figures 8 and 9. These graphs illustrate the performance of the three algorithms (CAMshift, MOTLD and the proposed method) over a whole tracking sequence. One can notice that the proposed method keeps track of the people face until the end for the three considered sequences, while CAMshift diverges at the 10^{th} frame for the front view sequence and at the 25^{th} frame for the 3/4 view sequence. MOTLD fares better on the front view, but diverges at the 19^{th} frame for the 3/4 view, and at the 35^{th} frame for the profile view. Figures 10, 11 and 12 illustrate the appearance of the divergence. One can also notice that in the 3/4 view sequence, after losing track of the target, CAMshift manages to re-converge on the tracked face (approximately at the 66^{th} frame). To summarize, in the three sequences, the tracking results obtained by the proposed approach are satisfactory and permits to anonymize correctly the user and to significantly reduce database annotation time (ie. 60ms per frame, fast enough for our purposes).

V. Conclusion

In this paper, a new intuitive and semi-automatic video annotation software specifically designed for the needs of the EVEREST project is proposed. This software allows to create a ground truth annotation for three types of hazardous user behaviour in guided mountain transport, and track user faces automatically so that they can be blurred for the dissemination of the anonymous database. More precisely, a new face tracking method based on the combination of CAMshift and

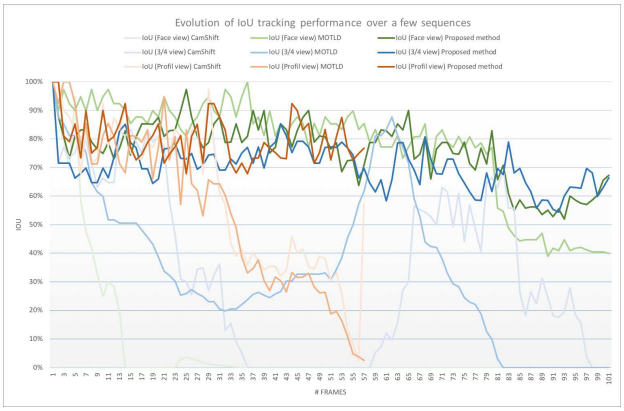

Fig. 8: Evolution of IoU tracking performance over a few sequences

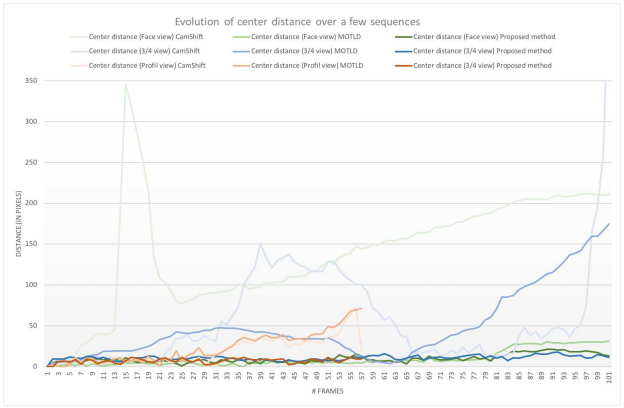

Fig. 9: Evolution of center distance over a few sequences

Farnebäck optical flow is proposed. The use of optical flow to guide CAMshift led to a more robust, stable, and accurate tracking algorithm. Indeed, for the three tested sequences, the proposed approach obtains an IoU score above 70%. In future works, we plan to test the robustness of the proposed approach on all sequences of the EVEREST video database including others view such as back and top view). We also consider extending the work presented in this paper to create even more robust algorithms based on CAMshift and optical flow to adverse conditions (weather and moving background).

Acknowledgment

This Work is being developed in the framework of EVEREST Project: a research activity that is financially supported by STRMTG (Service Techniques des Remontées Mécaniques et des Transports Guidés) of the French Ministry of Ecological and Solidarity Transition.

References

[1] Bascol K., Emonet R., Fromont E., Debusschere R. "Improving Chairlift Security with Deep Learning". In: Adams N., Tucker A., Weston D. (eds) Advances in Intelligent Data Analysis XVI. IDA 2017. Lecture Notes in Computer Science, vol 10584. Springer, Cham, pp. 1-13, 2017

[2] Gary R. Bradski, Microcomputer Research Lab, Santa Clara, CA, Intel Corporation, "Computer Vision Face Tracking For Use in a Perceptual User Interface" Intel Technology Journal, 2nd Quarter, 1998.

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Fig. 10: Evolution of tracking performance in front view over 100 frames (top to bottom: CAMshift, MOTLD and our method).

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Frame #1 Frame #5 Frame #10 Frame #25 Frame #50 Frame #100

Fig. 11: Evolution of tracking performance in 3/4 view over 100 frames (top to bottom: CAMshift, MOTLD and our method).

[3] Farnebäck, “Two-Frame Motion Estimation Based on Polynomial Expansion“, In: Bigun J., Gustavsson T. (eds) Image Analysis. SCIA 2003. Lecture Notes in Computer Science, vol 2749. Springer, Berlin, Heidelberg, pp. 363-370, 2003.

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

[5] Carl Vondrick, Donald Patterson, Deva Ramanan. ”Efficiently Scaling Up Crowdsourced Video Annotation” International Journal of Computer Vision (IJCV). June 2012.

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

[7] Bianco, Simone and Ciocca, Gianluigi and Napoletano, Paolo and Schettini, Raimondo, ”An interactive tool for manual, semi-automatic and automatic video annotation”, Computer Vision and Image Understanding, Elsevier, vol. 131, pp. 88-99, 2015

[8] Shen, Anting, ”BeaverDam: Video Annotation Tool for Computer Vision Training Labels”, EECS Department, University of California, Berkeley, Dec. 2016

[9] P.A. Viola, M.J. Jones, ”Rapid object detection using a boosted cascade of simple features”, CVPR, issue 1, pp. 511518, 2001.

[10] Redmon, Joseph and Farhadi, Ali, ”YOLOv3: An Incremental Improvement”, arXiv, 2018

[11] Liu, Wei and Anguelov, Dragomir and Erhan, Dumitru and Szegedy,

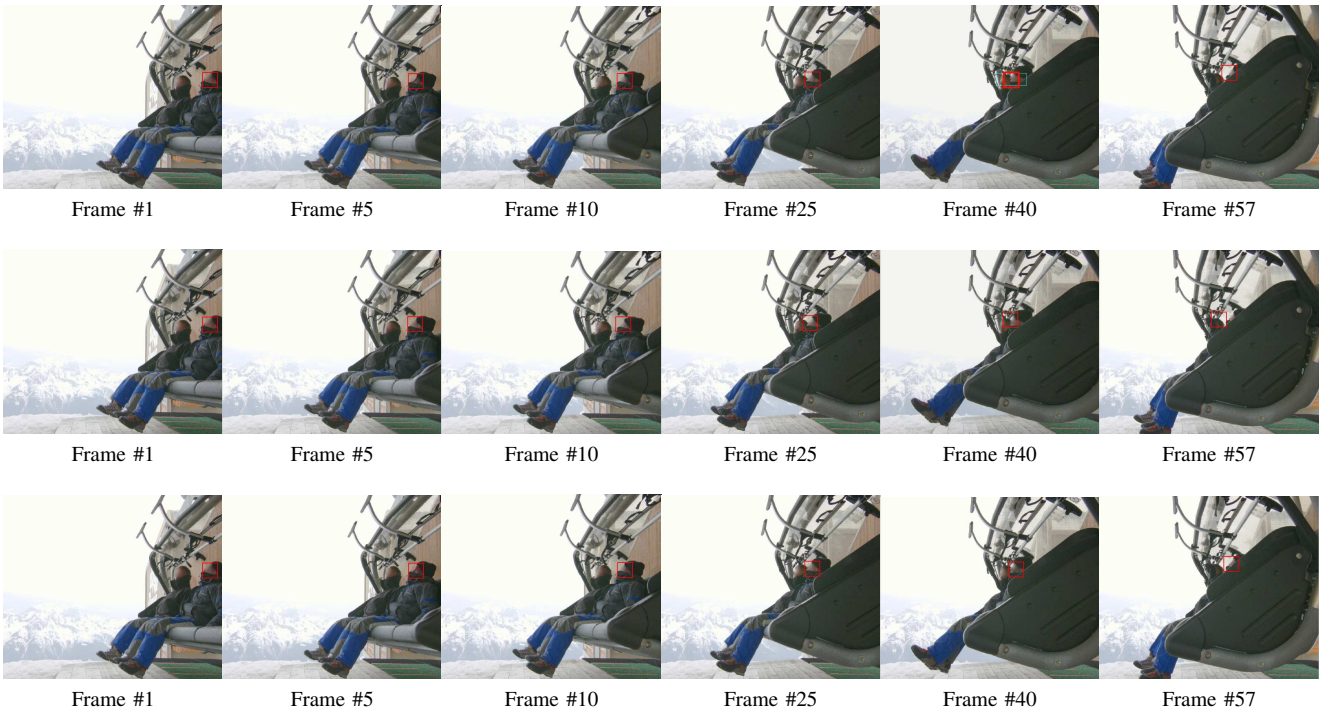

Fig. 12: Evolution of tracking performance in profile view over 57 frames (top to bottom: CAMshift, MOTLD and our method).

Christian and Reed, Scott and Fu, Cheng-Yang and Berg, Alexander C., "SSD: Single Shot MultiBox Detector", ECCV, 2016

[12] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering " IEEE Trans. Pattern Anal. Machine Intell., vol. 17, pp. 790-799, 1995.

[13] Gibson, J.J., "The Perception of the Visual World" (Riverside Press, Cambridge), 1950.

[14] Jianqin Yin, Yanbin Han, Jinping Li, Aizeng Cao, "Research on Real-Time Object Tracking by Improved Camshift", International symposium on Computer Network and Multimedia Technology, pp. 1-4, 2009

[15] Lucas, B., Kanade, T., "An Iterative Image Registration Technique with an Application to Stereo Vision" In: Proc. Darpa IU Workshop, pp. 121-130, 1981