Automatic video analysis framework for exposure region recognition in X-ray imaging automation

J. Sun, Z. Wu, Z Yu, H. Chen, C. Du, L. Xu, J. Zhong, J. J Feng, Gouenou Coatrieux, Jean-Louis Coatrieux, et al.

Automatic video analysis framework for exposure region recognition in X-ray imaging automation

Jiarui Sun, Zhan Wu, Zechen Yu, Huanji Chen, Changping Du, Liang Xu, Jian Zhong, Juan Feng, Gouenou Coatrieux, *Senior Member, IEEE*, Jean-Louis Coatrieux, *Life Fellow, IEEE*, and Yang Chen, *Senior Member, IEEE*

***Abstract*— The deep learning-based automatic recognition of the scanning or exposing region in medical imaging automation is a promising new technique, which can decrease the heavy workload of the radiographers, optimize imaging workflow and improve image quality. However, there is little related research and practice in X-ray imaging. In this paper, we focus on two key problems in X-ray imaging automation: automatic recognition of the exposure moment and the exposure region. Consequently, we propose an automatic video analysis framework based on the hybrid model, approaching real-time performance. The framework consists of three interdependent components: Body Structure Detection, Motion State Tracing, and Body Modeling. Body Structure Detection disassembles the patient to obtain the corresponding body keypoints and body Bboxes. Combining and analyzing the two different types of body structure representations is to obtain rich spatial location information about the patient body structure. Motion State Tracing focuses on the motion state analysis of the exposure region to recognize the appropriate exposure moment. The exposure region is calculated by Body Modeling when the exposure moment appears. A large-scale dataset for X-ray examination scene is built to validate the performance of the proposed method. Extensive experiments demonstrate the superiority of the proposed method in automatically recognizing the exposure moment and exposure region. This paradigm provides the first method that can enable automatically and accurately recognize the exposure region in X-ray imaging without the help of the radiographer.**

***Index Terms*— X-ray imaging automation, exposure region recognition, computer vision, video analysis, deep learning**

This work was supported in part by the State's Key Project of Research and Development Plan under Grant 2017YFC0109202 and Grant 2017YFA0104302, in part by the National Natural Science Foundation under 61871117, in part by Science and Technology Program of Guangdong (2018B030333001). (Corresponding author: Yang Chen).

J. Sun, Z. Yu, H. Chen, and C. Du are with the School of Computer Science and Technology, Southeast University, Nanjing, China, (e-mail: 230198566@seu.edu.cn, 220191708@seu.edu.cn, 513748165@qq.com, 220191716@seu.edu.cn).

Z. Wu is with the School of Cyberspace Security, Southeast University, Nanjing, Jiangsu, China (e-mail: zhanwubusheng1994@foxmail.com).

L. Xu, J. Zhong, and J. Feng are with the XR IC Department, United Imaging Healthcare, Shenyang, China (e-mail: skytiger00000@163.com, sau_zhongjian@163.com, 13352433355@163.com).

G. Coatrieux is with the IMT Atlantique, Inserm, LaTIM UMR1101, Brest 29000, France (e-mail: gouenou.coatrieux@telecom-bretagne.eu)

J.-L. Coatrieux is with the Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, F-35000 Rennes, France, with the Centre de Recherche en Information Biomédicale Sino-francais, 35042 Rennes, France, and also with the National Institute for Health and Medical Research, F-35000 Rennes, France (e-mail: jean-louis.coatrieux@univ-rennes1.fr)

Y. Chen is with the Lab of Image Science and Technology, Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 210096, China, and also with the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China (e-mail: chenyang.list@seu.edu.cn).

I. Introduction

DIGITAL radiography (DR) is one of the most affordable and frequently used medical imaging techniques, which has the advantages of easy accessibility and economy [1]. Because of convenience and low radiation dose in clinical routine compared to other imaging techniques, DR also is regarded as a preliminary screening method for some diseases [2], [3]. It can perform efficient body examinations, which greatly facilitates the diagnosis and treatment of the clinically serve and emergency patient [4], [5]. During an entire X-ray imaging, the radiographers assist the patient to pose, determine the exposure moment, identify the exposure region, adjust the X-ray collimator range, check and verify, and finally perform X-ray exposure. Usually, identifying the exposure region is a very important procedure in X-ray imaging, which is closely related to the radiation dose received by the patient. Moreover, the radiographer needs to determine the appropriate moment to successfully perform an exposure, which can avoid image retakes caused by the motion artifacts and the position errors [6], [7]. Avoiding image retakes is crucial to reducing unnecessary radiation dose and inconvenience of the patient, as well as in avoiding waste of medical resources for hospitals [8]–[10]. In X-ray imaging, the two key procedures including recognizing the exposure moment and the exposure region tend to require costly and error-prone manual involvement. It will be inspiring if the exposure moment and region can be automatically and accurately recognized without the help of the radiographer.

Deep learning is an effective method to improve the effectiveness and efficiency of clinical care in recent years [11],

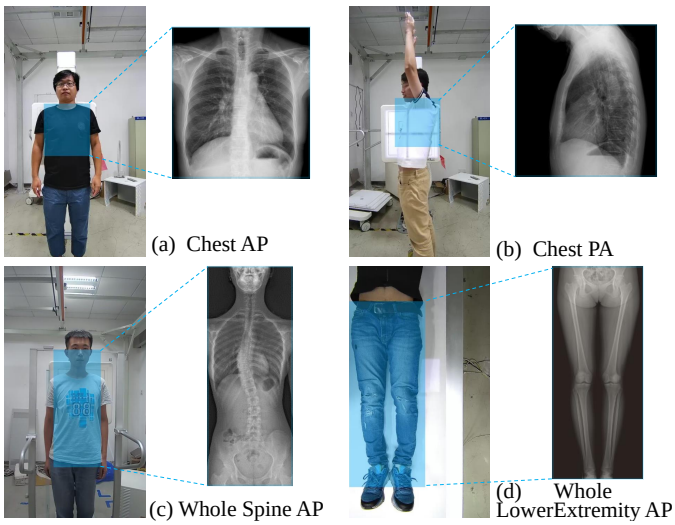

Fig. 1. Most exposure regions of X-ray imaging protocols usually consist of multiple or single incomplete body parts, and the exposure regions of different X-ray imaging protocols usually have larger overlapping areas. On given RGB images, the blue area indicates the standard exposure region. The area extended by the dotted line on the right is the corresponding X-ray radiographs.

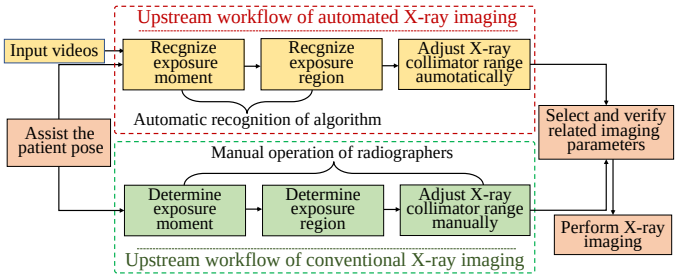

Fig. 2. For the exposure region in X-ray imaging, there is giving a comparison between conventional manual recognition and automatic recognition. In the upstream workflow of X-ray imaging, using automated methods will greatly reduce the workload of manual operation for the radiographer.

[12]. With the ever-increasing demand for health care services and the considerable drain on human resources, deep learning has infiltrated the optimization of clinical workflows [13], [14]. However, these optimizations almost usually focus on the downstream workflows, including the disease analysis and diagnosis [15], [16]. The upstream medical imaging workflows remain mostly unexplored [17]. Therefore, deep learning may facilitate automating the two key procedures in X-ray imaging, thereby decreasing the heavy manual workload of the radiographer, reducing the non-essential radiation dose for the patient, and optimizing imaging quality.

The deep learning-based automatic recognition of the scanning or exposing region in medical imaging automation has found relevant applications. U-HAPPY (United imaging Human Automatic PlanBbox for PulmonarY) is a successful attempt at computed tomography (CT) imaging using deep learning to automate pulmonary scanning [18]. It implements the automatic recognition of some scanning parameters during the pulmonary CT imaging. These parameters include the scanning region of standard pulmonary CT imaging and the moving distance of the scanning couch, etc. Another notable example is an automated scanning workflow based on the United imaging mobile CT platform [19]. Compared to U-HAPPY [18], it can automatically identify whether the patient is deemed ready using the motion analysis algorithm. Booijet et al. [20] and Saltybaeva et al. [21] have calculated the ISO-centering parameter in CT examination using the 3D camera algorithm. The ISO-centering parameter is calculated by the results of patient contour detection and guides the CT table automatically to adjust so that the center of the patient body region overlaps with the scanner ISO center. Kagan Incetan et al. [22] has developed a safety system based on an RGB-D camera to automate the patient positioning process of digital rotational angiography (DRA) imaging. The developed algorithm determines collisions between the C-arm and patient and the re-protocol algorithm identifies the movement of the patient table required to ensure a collision-free scan. Compared to scanning region recognition of CT or DRA in imaging automation, it is more challenging for the exposure region recognition in X-ray imaging. These challenges include: (1) X-ray examination involves more imaging protocols compared to CT or DRA, which means that the types of the exposure region that the algorithm needs to recognize are diverse. (2) The most exposure region usually consists of multiple (Fig.1 (c) and (d)) or single incomplete body parts (Fig.1 (a) and (b)). Thus, this causes difficulty in the area feature description and extraction. (3) Exposure regions of some special imaging protocols need to distinguish the left-right attribute. For single object detection methods, if the exposure region of every imaging protocol is set to different category Bboxes to conduct detection tasks, this is not technically feasible. This is because the object detection method cannot further recognize every specific instance in these detected Bboxes with the left-right attribute. (4) The exposure regions of different X-ray imaging protocols usually have larger overlapping areas. Current human parsing methods based on semantic segmentation [23], [24] usually can only define only one semantic category for each pixel position in the image. In this scene, because the semantics definition problem of overlapping areas in the exposure region cannot be solved, single semantic segmentation methods cannot handle it. To overcome the aforementioned challenge, we propose a robust framework based on the hybrid model to automatically recognize the exposure moment and region.

In this paper, we make the following contributions:

- For the first time, we contribute a near-real-time video analysis framework to automatically recognize the exposure moment and region during the X-ray imaging. The proposed framework has shown good recognition performance in experiments, which hopefully helps to decrease the radiographer workload and optimize X-ray imaging workflow.
- To the best of our knowledge, neither the single segmentation nor the object detection model can handle well the exposure region recognition of various X-ray imaging protocols. Therefore, we proposed a method based on the hybrid model including the body keypoint

TABLE I

12 KINDS OF X-RAY IMAGING PROTOCOLS FOR STANDING OR LYING EXAMINATION STATE IN OUR STUDY. AP MEANS ANTERIOR-POSTERIOR, PA MEANS POSTERIOR-ANTERIOR AND LAT MEANS LATERAL.

protocol	state		basic Bbox
	standing	lying	
Chest AP	✓	✓	B_{torso}
Chest PA	✓		B_{torso}
Chest LAT	✓	✓	B_{torso}
TSpine AP		✓	B_{torso}
TSpine LAT		✓	B_{torso}
LSpine AP		✓	B_{torso}
LSpine LAT		✓	B_{torso}
Whole Spine AP	✓	✓	B_{torso}
Whole Spine LAT	✓	✓	B_{torso}
Whole LowerExtremity AP	✓	✓	B_{person}
L-Whole LowerExtremity LAT	✓	✓	B_{person}
R-Whole LowerExtremity LAT	✓	✓	B_{person}

detection model and object detection model to calculate the exposure region.

- We constructed a large-scale dataset about the X-ray examination scene. The dataset includes a total of 6268 images which involves different 12 X-ray imaging protocols. Each image has corresponding annotations containing 9 categories of body bounding Bbox (Bbox) and 21 body keypoints. Currently, we are continuing to collect images about more protocols to enrich the dataset and test the robustness of the proposed framework. The dataset may promote the related clinical research. Soon, dataset[1] will be uploaded and made available for the researchers.

II. MOTIVATION AND PROBLEM STATEMENT

1) Motivation: As illustrated in Fig.2, in the upstream workflow of conventional X-ray imaging, radiographers require to artificially determine the exposue moment and region, and adjust the X-ray collimator range. Using automated methods instead of manual operation will greatly reduce the workload for the radiographers in the process. Therefore, we focus on the two key procedures in the upstream workflow of X-ray imaging: determining the exposure moment and recognizing the exposure region by an automated method without radiographer involvement.

2) Problem Statement: In this paper, we studied automatic recognition under 12 lying or standing X-ray imaging protocols (as listed in Table I). In these protocols, the exposure regions also are diverse, the patient body postures are varied during the examination. As illustrated in Fig.2, our goal is to automatically and accurately recognize the exposure moment and region without the involvement of the radiographers. To design a robust framework that can be hopefully employed in the clinical environment, several requirements should be satisfied:

(1) Preferably, the framework should not be invasive and introduce no extra radiation dose to the patient.
(2) In the X-ray examination, the motion state of the exposure region should be always tracked. The framework should have the approaching real-time performance so that the appropriate exposure moment can be found in time. Meanwhile, the approaching real-time performance also helps the radiographer to find abnormal situations and rapidly handle them.
(3) The accuracy of the exposure region recognition should be high. This avoids imaging failures or unnecessary radiation doses to the patient. Besides, the proposed framework also can be compatible with more protocols without making excessive modifications and sacrificing accuracy.

[1]https://github.com/JiaRuiS/AVAF

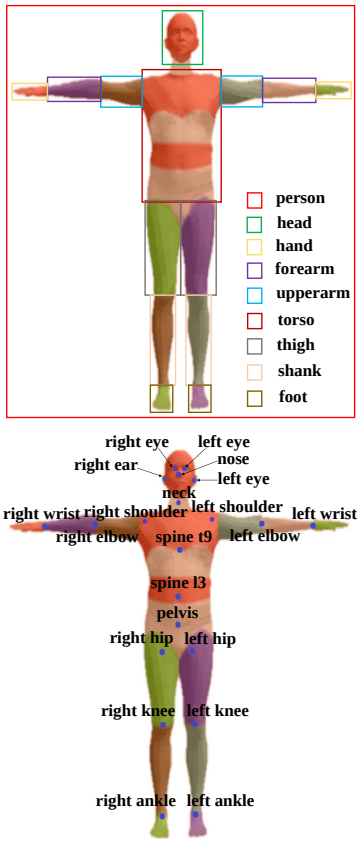

Fig. 3. The body Bbox annotations that need to be detected have 9 categories. The body keypoint annotations that need to be detected have 21 categories.

To meet the aforementioned requirements, we proposed the framework based on the hybrid model including body keypoint detection and body part detection models. Especially, when the framework runs, the information of the motion state or abnormal situations are marked on each video frame and fed back to the radiographer rapidly.

III. METHOD

The pipeline of our proposed framework is described in Fig.4. The framework consists of three interdependent components: (a) Body Structure Detection including Body Part Detection and Body Keypoint Detection is designed to obtain two complementary body structure representations for each patient. (b) Motion State Tracking is utilized to analyze the motion state of the exposure region, and recognize the exposure moment. (c) Body Modeling is used to model the

patient using the obtained two complementary body structure representations. Finally, the exposure region can be calculated directly.

A. Body Structure Detection

Body Structure Detection is designed to provide a basis for calculating the exposure region, which consists of Body Part Detection and Body Keypoint Detection. Body Part Detection detects the body Bboxes of the patient based on the object detection model. Body Keypoint Detection detects the body keypoints of the patient based on the keypoint detection model.

The border definition of the exposure region is clarified strictly by clinical norms. In the proposed method, the border of the exposure region will be represented by the body Bboxes and the body keypoints. Therefore, the number and category of the body Bboxes and the body keypoints will be determined by the categories of involved X-ray imaging protocols. In our current study, the body Bboxes are designed into 9 categories, and the body keypoints are designed into 21 categories.

1) Body Part Detection: As illustrated in Fig.3, for Body Part Detection, the body Bboxes of 9 categories will be detected: B_{person}, B_{head}, B_{torso}, $B_{upperarm}$, $B_{forearm}$, B_{hand}, B_{head}, B_{shank}, B_{foot}. The pair Bboxes (eg. left upperarm and right upperarm) should be classified further between left or right. This is because it can provide more necessary information to calculate the exposure region of these special imaging protocols including L-Whole LowerExtremity LAT and R-Whole LowerExtremity LAT. However, each instance in the pair Bboxes is not further identified. This is because object detection methods classify images in the form of area, lack the perceptual ability for position relationships between the Bboxes of different categories. Because the body keypoints contain detailed instance information, Body Keypoint Detection is introduced to help solve the above problem. Especially, B_{person} represents the Bbox of the people and is used to locate the patient during the X-ray imaging.

Detection model. In Body Part Detection, You Only Look Once version 5 (YOLOv5) [25] is employed to detect the body Bboxes of the aforementioned 9 categories. This choice was inspired by a comparison study (TABLE III) that comprehensively compared the performance of recent state-of-the-art (SOTA) object detection methods on the dataset. YOLOv5 follows the one-stage framework. It includes backbone part, neck part, and detection head part. Backbone part extracts multiscale features from the input image using fused Focus and CSP [26] structure. Neck part strengthens the integration and utilization of semantic features of different levels using FPN [27] and PAN [28] methods. Detection head part performs the final classification and regression to locate each body part. To match detection categories for our task, the output size of the detection head is adjusted to (80, 80, 52), (40, 40, 52), (20, 20, 52). In model training and testing, the size of all input images is resized to 640×640.

2) Body Keypoint Detection: As illustrated in Fig.3, for Body Keypoint Detection, the body keypoints of the 21 categories will be detected: P_{nose}, $P_{lefteye}$, $P_{righteye}$, $P_{leftear}$, $P_{rightear}$, $P_{leftshoulder}$, $P_{rightshoulder}$, $P_{leftelbow}$, $P_{rightelbow}$, $P_{leftwrist}$, $P_{rightwrist}$, $P_{lefthip}$, $P_{righthip}$, $P_{leftknee}$, $P_{rightknee}$, $P_{leftankle}$, $P_{rightankle}$, P_{neck}, P_{t9}, P_{l3}, P_{pelvis}. Compared to the body Bboxes, the body keypoint contains more detailed instance information (eg. lefteye and righteye). This is because the training and testing of pose estimation method is based on the whole image. Therefore, the spatial position relationship between different body keypoints can be well obtained.

Detection model. In Body Keypoint Detection, Alpha pose [29] is utilized to detect the body keypoints of the aforementioned 21 categories. This choice was inspired by a comparison study (TABLE III) that comprehensively compared the performance of several state-of-the-art pose estimation methods on the dataset. Alpha pose follows the top-down framework. VGG-based SSD-512 [30] detects the human Bboxes. Stacked Hourglass model [31] with Symmetric Spatial Transformer Network (SSTN) generates body keypoints for each given human Bbox. Parametric Pose NMS (p-Pose NMS) eliminates redundancy for the results of generated body keypoints to obtain fined body keypoints. In this paper, to improve computational efficiency, Body Keypoint Detection only calculates body keypoints for the patient. The human body Bbox of the patient is obtained from B_{person} predicted by Body Part Detection. For Stacked Hourglass model, we used a smaller 4-stack hourglass network and adjusted the number of predicted heatmap channels to 21 to adapt to our task. In model training and testing, the size of all input images is resized to 256×192.

B. Motion State Tracking

Motion State Tracking analyzes the motion state of the exposure region and recognizes the appropriate exposure moment during the X-ray imaging. As illustrated in Fig.4, Motion State Tracking is performed in the form of several sub-processes. Before the X-ray examination starts, the corresponding imaging protocol will be selected by the radiographer. It will tell the framework the category of the exposure region to be recognized.

As illustrated in Fig.4, Body Part Detection is performed firstly when Body Part Detection starts. In analysis 1, for each given video frame, the number of the B_{person} will be counted. Thus, the number of people that have entered the area within the field view of the camera can be determined. Body Part Detection is only performed in Motion State Tracking stage. Therefore, the exposure region cannot be calculated only using body Bboxes. To analyze the motion state of the exposure region, the basic body Bbox is set as an approximate substitute to represent the exposure region. As listed in Table I, the basic body Bbox is the category that has the largest intersection over union (IoU) with the exposure region of the given X-ray imaging protocol. For a given frame, $B_{basic} = \{b_1, b_2, ..., b_i\}$ represents the detection results for all basic Bboxes. The position information of Flat Panel Radiation Detector (FPRD) can be captured in real-time from the DR system. Thus, the additional computational overhead for detecting the position of FPRD can be avoided. The position of FPRD is defined in the form of the smallest body Bbox containing the FPRD and is set as ROI_{FPRD}. Then, p is formulated to measure the

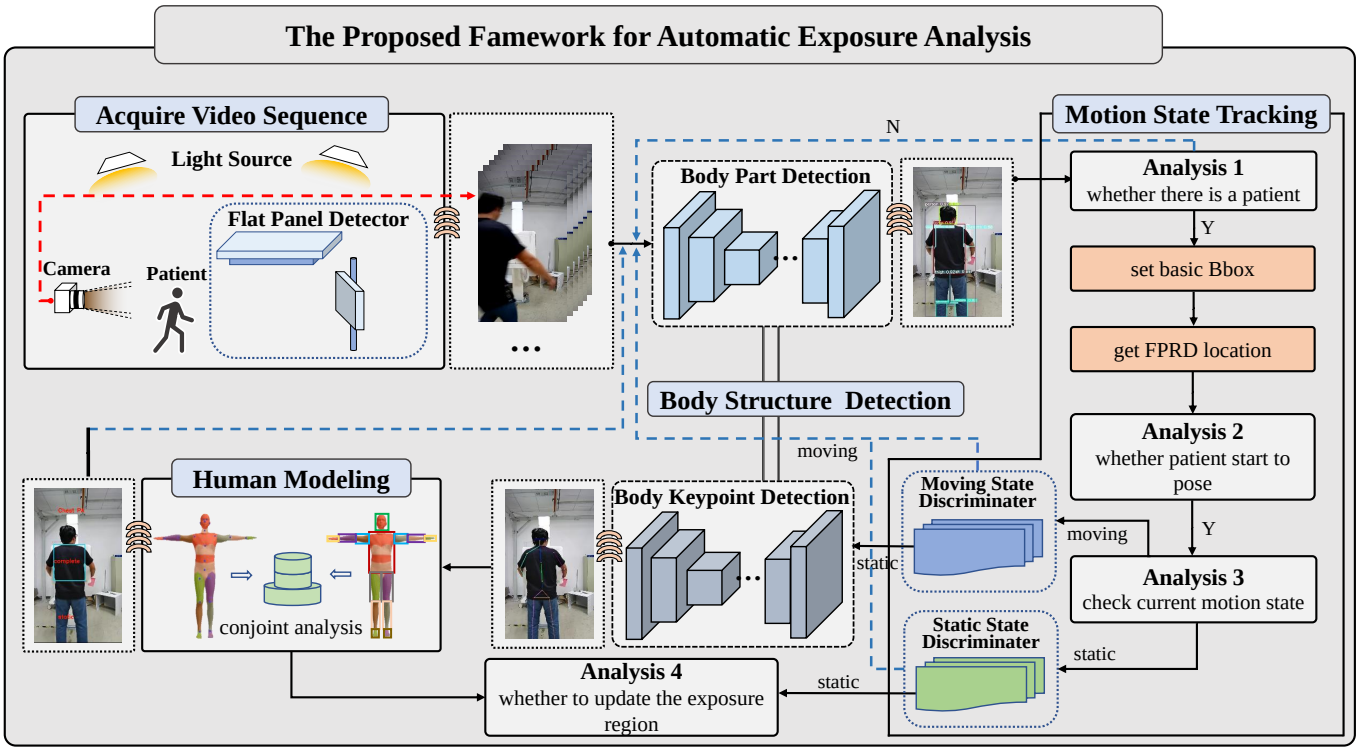

Fig. 4. The pipeline of the proposed framework for automatic recognition of the exposure moment and region.

spatial position closeness of b_i and ROI_{FPRD}:

$$p = \frac{|b_i \cap ROI_{FPRD}|}{|ROI_{FPRD}|}, \tag{1}$$

where $|\cdot|$ is the number of pixels within the given area. T_r is the threshold to measure whether ROI_{FPRD} and b_i are close enough in the spatial position. When $p > T_r$, it indicates that the given b_i and ROI_{FPRD} are close enough. Under the circumstance, the patient is ready to start posing. However, B_{basic} possibly includes multiple elements. This is because sometimes the family member will assist the patient to pose. Meanwhile, there may be multiple b_i that satisfies $p > T_r$. Actually, there are existing only b_i within ROI_{FPRD} when the assistance process ends. Thus, this does not affect recognizing the exposure moment.

Next, the motion state of ROI_{FPRD} will be used to analyze to recognize the exposure moment. This is because ROI_{FPRD} will completely contain the exposure region, and the exposure region of the current protocol stays in a static state. Therefore, the motion states of b_i and ROI_{FPRD} stay synchronized from starting posing to completing X-ray exposure. In this process, $N = \{M, S\}$ is used to describe the motion state of ROI_{FPRD}. Especially, the analysis of the motion state is based on the computation for successive frames. $N_f = \{m, s\}$ is used to describe the motion state of the single frame. For a given frame, s indicates the ROI_{FPRD} stays in static, and m indicates ROI_{FPRD} stays in motional. Because N always changes from M to S in the examination process, N is initialized to M when the imaging starts. In analysis 3, N is checked first:

- When $N = M$, Static State Discriminator is employed to identify whether N conducts the conversion from M to S. The dense optical flow method [32] is utilized to capture the displacement field between continuous two frames. To pursue real-time performance, each frame is down-sampled before the displacement field calculation. The matrix group $D = \{d_x, d_y\}$ are the displacement fields in the horizontal and vertical directions of ROI_{FPRD}. o describes motion intensity of the displacement field and is formulated as:

$$o = \frac{\sum_{i,j} (d_x(i,j) + d_y(i,j))}{|ROI_{FPRD}| \cdot t}, \tag{2}$$

where i and j represent the $(i,j)^{th}$ pixel in ROI_{FPRD}, and t is the discriminating factor that determines whether the single pixel of ROI_{FPRD} has a certain degree of displacement. To obtain the motion state of the given frame, threshold T_s is used to measure the motion intensity. When $o < T_s$, $N_f = s$. Otherwise, $N_f = m$. The criterion of state conversion from M to S is based on successive frames. Therefore, N is converted from M to S if N_f is s in continuous k frames. In our previous experiment, we found that sometimes some tiny image changes between continuous frames caused wrong results of Body Part Detection. Therefore, state-refresh mechanism is designed to avoid continuous impact from detection failure. With the state-refresh mechanism, the conversion relation of Static State Discriminator is formulated as:

$$N = \begin{cases} M, & n < k \\ S, & n = k \\ M, & n > r \end{cases}, \tag{3}$$

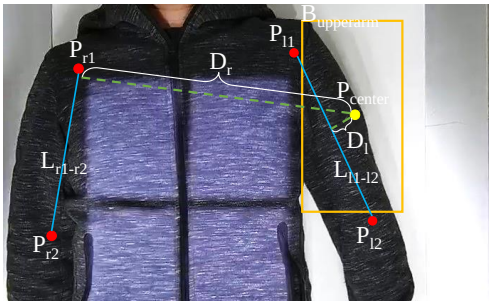

Fig. 6. The left-rightness information of the body Bboxes can be obtained by combining the body keypoints

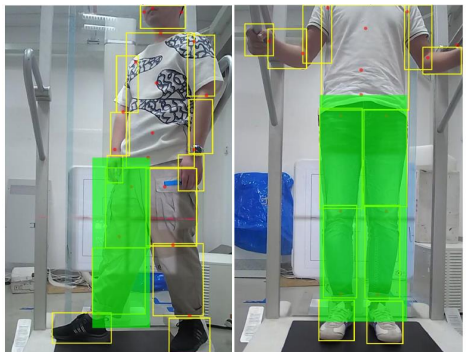

Fig. 7. Every border of the exposure region is obtained directly from the body Bboxes or the body keypoints, or using simple computation after Body Modeling. The green area indicates the calculated exposure region for the given X-ray imaging protocol.

$$h = \begin{cases} h_1, & -45^\circ \leqslant \alpha < 45^\circ \\ h_2, & 45^\circ \leqslant \alpha < 135^\circ \\ h_3, & 135^\circ \leqslant \alpha < 225^\circ \\ h_4, & 225^\circ \leqslant \alpha < 315^\circ \end{cases}, \tag{9}$$

where h_1 is the direction h generally toward the positive half-axis of the x-axis in the plane coordinate system. The rest direction can be deduced by analogy. Therefore, the directionality information of the body Bboxes can be obtained.

The distance between center point of the body Bbox and the line of the possible adjacent two body keypoints is computed to obtain the left-rightness information of the body Bboxes. The keypoint group is defined as $P_g = \{P_{l1}, P_{l2}, P_{r1}, P_{r2}\}$, which means the two pairs of body keypoints that are adjacent the given body Bbox. P_{l1} and P_{l2} are the pair of the body keypoints on the left side of the patient, and P_{r1} and P_{r2} are another pair on the right side. Each body Bbox with the left-rightness can correspond to the only keypoint group. As illustrated in Fig.6, P_{center} is the center coordinate of the given Bbox. D_l is the distance from point P_{center} to line L_{l1-l2} connected by P_{l1} and P_{l2}. D_r is the distance from point P_{center} to line L_{r1-r2} connected by P_{r1} and P_{r2}. Directionality of the each body Bbox is calculated according to the size relationship between D_l and D_r: the body Bboxes of left-rightness is left when $D_l \leqslant D_r$. Otherwise, left-rightness is right. Therefore, the left-rightness information of the body Bboxes can be obtained.

Meanwhile, the directionality and left-rightness of the body Bboxes are determined. As presented in Fig.7, for the given X-ray imaging protocol, each one in the four borders (top, down, left, and right) of the exposure region can be obtained directly from the body Bboxes or the body keypoints, or using simple computation.

IV. EXPERIMENTAL RESULTS

A. Construction of Datasets

All experiment data were provided by the company of United Imaging Healthcare (UIH) to validate the effectiveness of the proposed method. These data were collected by simulating the X-ray imaging process, but real exposure was not performed. A large-scale dataset called the X-ray examination scene dataset (XES) was constructed after data collection and processing. The study (data collection and processing) was approved by the institutional review board of UIH and was adherent to the tenets of the Declaration of Helsinki.

Data collection. Raw data were obtained in video form. And raw videos were collected on patient-level: each patient simulated X-ray imaging processes of different protocols and generated a corresponding video. Thus, the content of each raw video only involves a single patient. These videos were collected from 135 patients including 89 males and 46 females. Among them, the age of patients ranged from 20 to 55 years old, with an average age of 34 years. To ensure the algorithm robustness, the collected data is diverse in the patients and the collection environment. Therefore, the patients have different dresses, body sizes, skin colors, and genders in the XES. The environment diversity includes different examination rooms, light intensities, source to image receptor distance (SID) (SIDs are usually diverse between different protocols) in the XES. To build a high-quality dataset, all high-resolution videos were collected using the 2D camera (HIKVISION, DS-2CD6424FWD-C1: 50Hz, 25fps, 1280 $\times$ 720 pixes) installed on the uDR-WuKong. All collected videos are RGB modal and obtained from September 2020 to November 2020.

Data processing. The constructed dataset consists of two parts: Video part and Image part. For Video Part or Image Part, the data split between Training-Validation (TV) set and Test set is first on patient-level. After the split on patient-level, each raw video in Video Part is divided into several sub-videos with the X-ray imaging protocol as the base unit. Thus, this ensures that there is no data overlap between the TV set and Test set. Video part is divided into two subsets: TV set (300 sub-videos) and Test set (392 sub-videos). Because the Body Part Detection and the Body Keypoint Detection are based on single image, the TV set of Video part is used to generate Image part. The Test set of Video part is used to test the entire performance of the proposed framework. In the process of generating Image Part, each sub-video is sampled at equal intervals. The adjacent frames in the sub-videos usually play the same contribution to the model training because they have almost no difference in the image. Therefore, the similarity between frames evaluated by Hamming Distance [33] is calculated to remove overmuch adjacent frames after sampling to obtain Image Part. The two subsets of Image part serve the training and testing of two different detection

TABLE II
STATISTICS DISTRIBUTION FROM THE XES.

Positioning	Video Part		Image Part	
	TV set (S/L)	Test set (S/L)	TV set (S/L)	Test set (S/L)
Chest AP	10/29	11/40	140/463	51/168
Chest PA	9/-	10/-	129/-	47/-
Chest LAT	11/24	12/33	148/389	54/141
TSpine AP	-/10	-/14	-/187	-/68
TSpine LAT	-/12	-/16	-/167	-/61
LSpine AP	-/9	-/12	-/159	-/58
LSpine LAT	-/10	-/13	-/144	-/52
Whole Spine AP	13/23	15/32	143/375	52/136
Whole Spine LAT	10/26	11/36	121/404	44/146
Whole LowerExtremity AP	16/29	19/40	206/551	75/199
L-Whole LowerExtremity LAT	11/19	13/26	130/315	47/114
R-Whole LowerExtremity LAT	8/21	10/29	107/322	39/116
Total	88/212	101/291	1124/3476	409/1259

models respectively. The first subset (TV set) is employed for training and validation. The second subset (Test set) is used for independent testing. As shown in Fig.3, each image of Image Part has these annotations containing the body Bboxes of 9 categories and the body keypoints of 21 categories. All the labels are annotated strictly by 10 experienced radiographers. Then, a senior imaging expert with several years of experience in X-ray imaging performed quality control on the annotated dataset. The detailed statistics of the XES can be seen in Table II. Especially, previous experiment results about the detection models show: the model detection ability under the standing imaging protocols is more capable than the lying imaging protocols when using the same amount of training data. Therefore, the extra data under the lying imaging protocols are collected and supplemented to the XES.

B. Implementation details

The proposed framework is implemented and runs in the following configured computer platform: CPU is Inter(R) Core(TM) i5-8500K 3.00GHz, and GPU is NVIDIA GTX-1660 super with 6G memory. The CUDA version is 10.0. Due to the friendliness of PyTorch 1.5.1 [34], we first employed it to respectively train the models of Body Part Detection and Body Keypoint Detection on a single GPU, which the adam optimizer [35] is utilized to optimize the two networks separately in the training process. Then, the corresponding version of LibTorch is used to convert the two trained models into the codes of the C++ version to integrate. The resolution of all input images is 720×1280.

Evaluation metrics. Quantitative and qualitative performance evaluations of the proposed method are given in the next two sub-sections. VOC2007 [36] and MSCOCO [37] metrics are employed to evaluate the model of Body Part Detection. Average precision (AP) indicates the detection precision of each category Bboxes, while mean average precision (mAP) shows the overall performance of all category Bboxes. Two widespread IoU thresholds (0.5 and 0.75) are utilized to obtain the corresponding mAP. The mAP(@0.5:0.95) indicates the mean performance of mAP under different IoU thresholds, in which the calculation interval between the adjacent two mAPs is 0.05. Besides, mean precision (MP) and mean recall (MR) also are utilized to evaluate the classification performance. The MP and MR are defined by Eqs.(10)-(11):

$$MP = \frac{1}{c}\sum_{i=1}^{c}\frac{TP}{TP+FP}, \tag{10}$$

$$MR = \frac{1}{c}\sum_{i=1}^{c}\frac{TP}{TP+FN}, \tag{11}$$

where TP is the number of the correct detected samples. FP is the number of erroneous detected samples. FN is the number of undetected ground truth samples. In the performance evaluation of Body Part Detection, the correct detected body Bbox samples satisfy following two conditions: (1) correct classification and (2) IoU $\geqslant 0.5$. In the performance evaluation of Body Keypoint Detection, the discrimination method of the correct detected keypoint samples is consistent with the PCK metrics in FLIC [38].

C. Result of Body Structure Detection

The proposed framework is based on the detection models of the body part and body keypoint. To determine the actual model in Body Part Detection and Body Keypoint Detection, extensive comparison experiments are conducted. The method with the best performance will be employed. The best results are retained to achieve the performance of each method after enough parameter adjustment experiments.

The performance of four recent SOTA methods of object detection is compared in Table III, including YOLO V3 [39], YOLO V5 [25], EfficientDet D0, and D1 [40]. All models are pre-trained on the object detection dataset of MSCOCO to obtain stronger feature extraction capabilities. Due to the limitation of the computation resources, EfficientDet of only D0 and D1 versions are evaluated. For all given methods, YOLO V5 achieves the best performance in each metric. As listed in Table IV, the detailed performance is evaluated for each category body Bbox using YOLO V5. Corresponding visual results also are given in the first line of Fig.8. Each category body Bbox can be detected well by YOLO V5. The detection of small objects (eg. hand) also shows a significant effect. This depends on the high-quality XES and the good method. Therefore, YOLO V5 is employed as the actual model of Body Part Detection.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS

Method	Body Part Detection					Keypoint Detection	
	MP	MR	mAP_{50}	mAP_{75}	mAP(@0.5:0.95)	MP	MR
EfficientDet D0	82.5	76.2	85.6	65.4	57.9		
EfficientDet D1	87.4	83.8	87.8	67.9	60.3		
YOLO V3	97.3	79.1	87.1	79.4	78.6		
YOLO V5	**97.8**	**92.7**	**99.6**	**89.2**	**85.2**		
HRNet						89.2	86.6
Alpha pose						**95.6**	**92.3**

TABLE IV
DETECTION PERFORMANCE FOR EVERY CATEGORYTHE FROM YOLO V5

	Person	Head	Torso	Upperarm	Forearm	Hand	Thigh	Shank	Foot	mean
Precision	99.6	99.1	96.8	97.2	96.2	96.8	98.5	98.4	98.4	97.8
Recall	97.8	91.2	93.7	85.8	87.2	92.6	97.7	94.7	94.0	92.7
AP_{50}	99.8	99.8	99.5	99.1	98.9	99.4	99.9	99.8	99.8	99.6

TABLE V
DETECTION PERFORMANCE FOR EVERY CATEGORYTHE FROM ALPHA POSE

	Nose	Eye	Ear	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Neck	T9	L3	Pelvis	mean
Precision	97.7	97.0	98.3	98.0	92.2	96.0	98.9	81.2	74.4	99.5	99.9	99.8	99.9	95.6
Recall	98.3	99.1	95.2	88.5	84.5	79.8	96.7	92.8	93.9	89.8	93.2	98.9	98.7	92.3

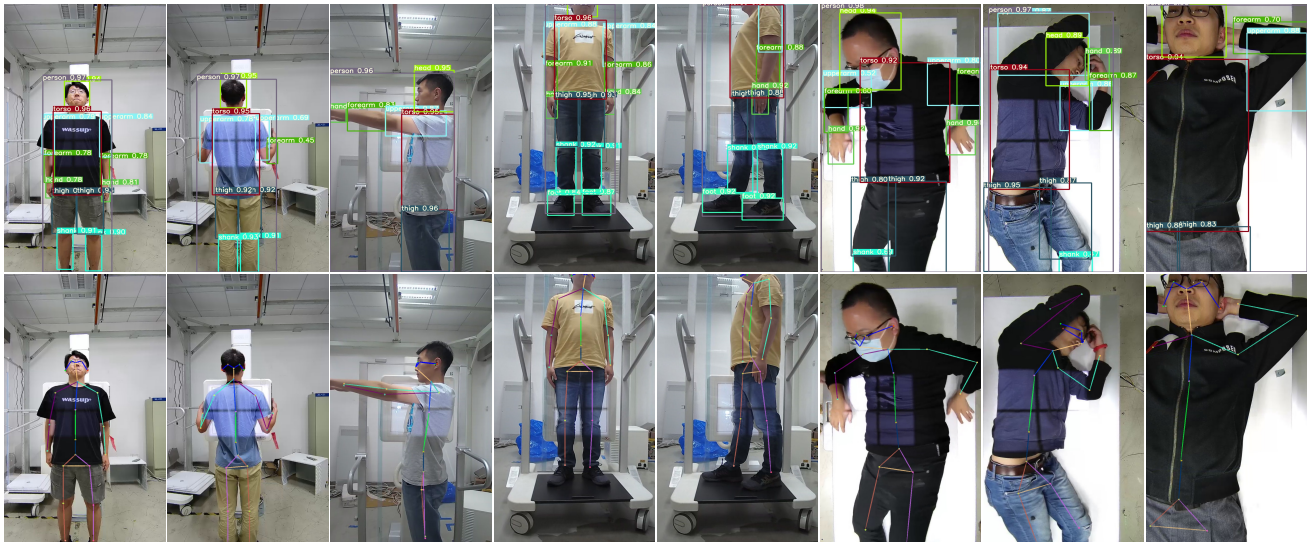

Fig. 8. Visual results of Body Part Detection and Body Keypoint Detection: the first line indicates the results of Body Part Detection, and the second line indicates the results of Body Keypoint Detection.

The performance of two recent SOTA methods of pose recognition is compared in Table III, including HRNet [41] and Alpha pose [29]. All models are pre-trained on the body keypoint detection dataset of MSCOCO. For all given methods, Alpha pose achieves the best performance in each metric. The patient Bboxes in the XES are utilized to locate the patient during the model training. The patient Bboxes are provided by Body Part Detection when inferring. As listed in Table V, precision and recall of each category keypoint are given using Alpha pose. Visual results of Body Keypoint Detection also are shown in the second line of Fig.8. In our previous experiments, each keypoint in pair (eg. left eye and right eye) usually shows similar performance. Therefore, only their average performance is given in Table V. For all given methods, Alpha pose achieves the best performance in MP and MR. Therefore, Alpha pose is employed as the actual model of Body Keypoint Detection. A lot of research and practice show the performance of pose recognition is closely related to the human body occlusion and integrity in the image [29], [42]–[45]. The SID is very small in some lying imaging protocols, which causes the failure for detecting some body parts. Therefore, the detection performance of the keypoints on the body edge (eg. hip and ankle) is usually not as good as other keypoints on the body center (eg. T9 and L3).

D. Result of Video Detection

As listed in Tables VI and VII, the performance of the proposed framework for the recognition of the exposure moment and region are evaluated. The evaluation is carried out by three radiographers with extensive clinical experience.

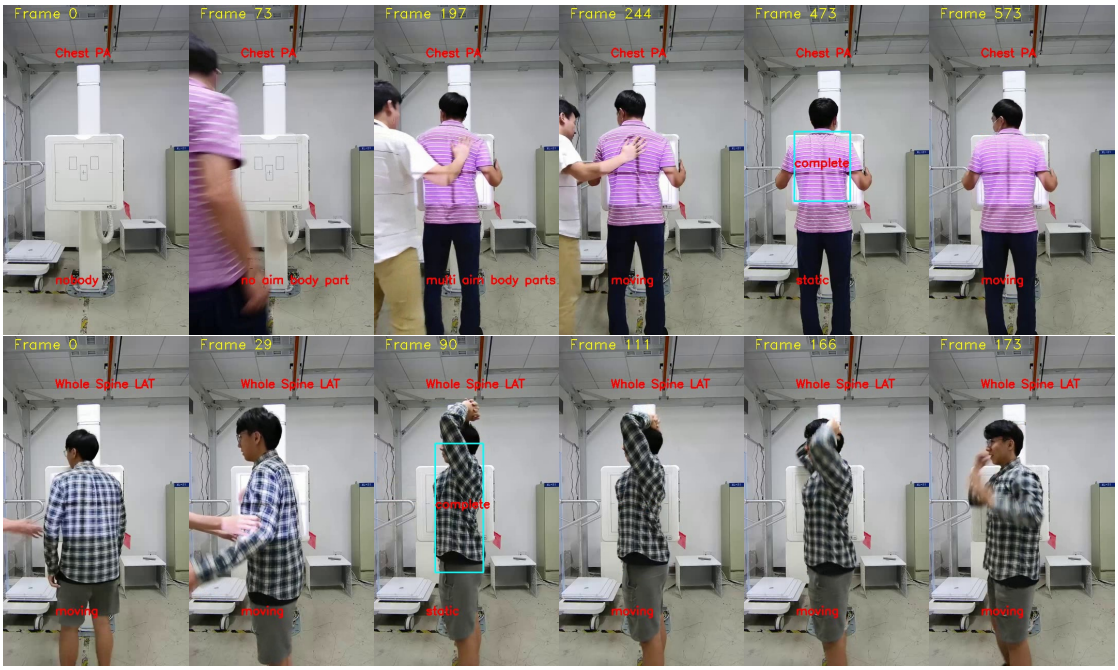

Fig. 9. Visual detection results for the standing X-ray imaging protocols. In each frame, the motion or abnormal states of the patient is marked in red font. The recognized exposure region is indicated in the blue Bboxes.

TABLE VI
VIDEO DETECTION PERFORMANCE FOR THE EXPOSURE MOMENT AND REGION IN THE STANDING X-RAY IMAGING PROTOCOLS

	Exposure moment	Exposure region
Precision	97.6	98.1
Recall	95.3	96.2

TABLE VII
VIDEO DETECTION PERFORMANCE FOR THE EXPOSURE MOMENT AND REGION IN THE LYING X-RAY IMAGING PROTOCOLS

	Exposure moment	Exposure region
Precision	95.0	91.1
Recall	91.0	88.9

Then, two senior X-ray imaging experts with rich clinical experience performed verification for evaluation. Precision and recall were employed to evaluate the entire performance of the proposed framework for detecting exposure moment and region. Specifically, the evaluation of the exposure region occurs only when the exposure moment appears. This is because the framework will calculate the exposure region only when the exposure moment appears. As can be seen in Tables VI and VII, the proposed framework has achieved an encouraging recognition performance for the standing or lying imaging protocols. Besides, the framework is executed on a GTX-1660 super GPU, which can achieve a frame rate of 10-14 fps. This means the proposed framework can be in near real-time only using the low-cost GPU.

As shown in Fig.9 and Fig.10, the visual results of the proposed framework are given by detecting the different sub-videos including the standing or lying imaging protocols. More visual results of video detection are provided in `https://github.com/JiaRuiS/AVAF`. As can be seen in Figs that the exposure moment and region can be recognized accurately under different X-ray imaging protocols. As described in METHOD section, the motion state of the FPRD area in each frame is divided into static or moving. Besides, the framework also warns abnormal states of the patient. According to Motion State Analysis, the framework gives the abnormal states that may occur during X-ray examination can be classified. These abnormal states follow: a) Nobody state indicates that the patient has not yet entered the field view of the camera in the X-ray examination room. b) No aim body part state indicates that the patient has not entered the imaging area of FPRD. c) Multi-aim body parts state indicates that multiple patients have entered the imaging area of FPRD. Besides, complete state indicates that the appropriate exposure moment has appeared and the exposure region can be calculated.

E. Discussion

The proposed framework has solved the two key problems in X-ray imaging automation, and the overall recognition performance of the exposure moment and region is encouraging. Besides, the proposed framework can provide timely feedback on abnormal conditions to radiographers. This can improve the work efficiency of the radiographers using guiding the patient to perform corrective actions in the current imaging process.

However, some certain limitations still need to be improved. Tables VI and VII present that although the proposed method has shown an encouraging performance under different X-ray imaging protocols. But the detection performance for the lying imaging protocols is not as good as the standing imaging protocols. The main reasons follow: (1) The exposure moment recognition involves the motion state analysis for the patient. The analysis process utilizes the optical flow method, and the method is sensitive to threshold selection. During

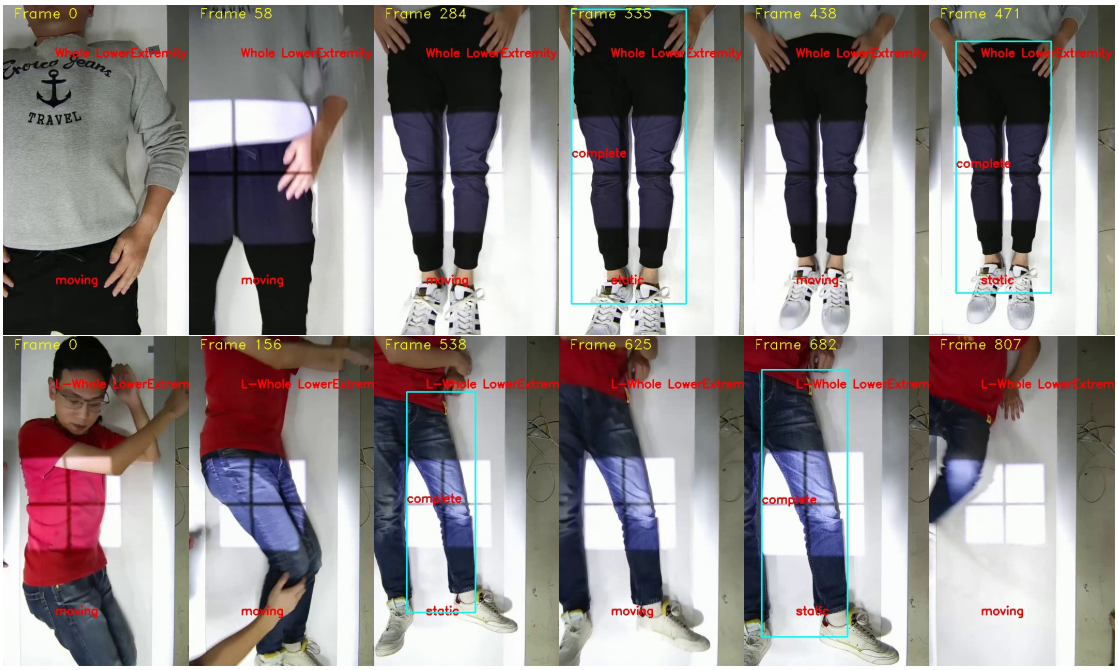

Fig. 10. Visual detection results for the lying X-ray imaging protocols. In each frame, the motion or abnormal states of the patient is marked in red font. The recognized exposure region is indicated in the blue Bboxes.

the examination, different radiographers may use different moving speeds to adjust the camera position and SID, so there is no guarantee that the optical flow method can work well in all different lying imaging protocols. (2) In standing imaging protocols, the SID usually is fixed. Therefore, the difference in image scale between different frames is smaller. For the lying protocol, the frequent changes of the SID cause a considerable scale diversity between frames. As we all know, the scale diversity has always been a challenging problem in computer vision [46]–[48]. Therefore, this brings a certain degree of difficulty to the detection of the body keypoints or body Bboxes. It may be lead to the wrong calculation for the exposure region when an error occurs in Body Part Detection or Body Keypoint Detection.

In our study, we also have tried to explore exposure region recognition about X-ray imaging protocols of some single body parts (only a single body part of the patient appears in the images). Because of the detection performance limitation of the keypoint detection model under such conditions, we have not been able to carry out further research on these imaging protocols. In the future, we will focus on the study of these protocols, and make them compatible with the existing framework.

V. Conclusion

In the paper, we propose a near-real-time video analysis framework to solve two key problems in X-ray imaging automation: the automatic recognition of exposure moment and exposure region. The framework includes three interdependent components: Body Structure Detection, Motion State Tracing, and Body Modeling. First, Body Structure Detection detects the body keypoints and the body Bboxes of the patient. The two different types of body structure representations are combined to obtain more rich spatial location information about the body structure. Second, Motion State Tracing analyzes the motion state of the exposure region to recognize the appropriate exposure moment. Finally, the exposure region is calculated by Body Modeling when the exposure moment appears. Extensive experiments demonstrate the superiority of the proposed method in the automatic recognition of exposure moment and exposure region. Besides, the framework can also track the motion state of the exposure region, analyze abnormal situations timely, and feed these information back to the radiographers. Therefore, the proposed framework is encouraging that it facilitates decreasing the radiographer workload and optimizing the upstream workflow in conventional X-ray imaging.

In future work, we will focus on the entire process of X-ray imaging automation from automatic recognition of exposure region to automatically realizing exposing for the patients, and validate the influence of the method by comparing the impact of manual imaging and automatic imaging on imaging quality.

References

[1] M. Korner, C. H. Weber, S. Wirth, K.-J. Pfeifer, M. F. Reiser, and M. Treitl, "Advances in digital radiography: physical principles and system overview," *Radiographics*, vol. 27, no. 3, pp. 675–686, 2007.

[2] M. Yaffe and J. Rowlands, "X-ray detectors for digital radiography," *Physics in Medicine & Biology*, vol. 42, no. 1, p. 1, 1997.

[3] C.-H. Lin, J.-X. Wu, C.-M. Li, P.-Y. Chen, N.-S. Pai, and Y.-C. Kuo, "Enhancement of chest x-ray images to improve screening accuracy rate using iterated function system and multilayer fractional-order machine learning classifier," *IEEE Photonics Journal*, vol. 12, no. 4, pp. 1–18, 2020.

[4] P. F. van der Stelt, "Better imaging: the advantages of digital radiography," *The Journal of the American Dental Association*, vol. 139, pp. S7–S13, 2008.

[5] S. Mc Fadden, T. Roding, G. De Vries, M. Benwell, H. Bijwaard, and J. Scheurleer, "Digital imaging and radiographic practise in diagnostic radiography: an overview of current knowledge and practice in europe," *Radiography*, vol. 24, no. 2, pp. 137–141, 2018.

[6] C.-S. Lin, P.-C. Chan, K.-H. Huang, C.-F. Lu, Y.-F. Chen, and Y.-O. Lin Chen, "Guidelines for reducing image retakes of general digital radiography," *Advances in Mechanical Engineering*, vol. 8, no. 4, p. 1687814016644127, 2016.

[7] M. M. Alipio and G. M. A. Lantajo, "Determinants of image retakes in general digital radiography," *Mindanao Journal of Science and Technology*, vol. 19, no. 1, 2021.

[8] D. Waaler and B. Hofmann, "Image rejects/retakes—radiographic challenges," *Radiation protection dosimetry*, vol. 139, no. 1-3, pp. 375–379, 2010.

[9] A. K. Jones, R. Polman, C. E. Willis, and S. J. Shepard, "One year's results from a server-based system for performing reject analysis and exposure analysis in computed radiography," *Journal of digital imaging*, vol. 24, no. 2, pp. 243–255, 2011.

[10] E. Vano, J. I. Ten, J. M. Fernandez-Soto, and R. M. Sanchez-Casanueva, "Experience with patient dosimetry and quality control online for diagnostic and interventional radiology using dicom services," *American Journal of Roentgenology*, vol. 200, no. 4, pp. 783–790, 2013.

[11] T. M. Maddox, J. S. Rumsfeld, and P. R. Payne, "Questions for artificial intelligence in health care," *Jama*, vol. 321, no. 1, pp. 31–32, 2019.

[12] S. J. Sirintrapun and A. M. Lopez, "Telemedicine in cancer care," *American Society of Clinical Oncology Educational Book*, vol. 38, pp. 540–545, 2018.

[13] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[14] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[15] Z. Wu, R. Ge, M. Wen, G. Liu, Y. Chen, P. Zhang, X. He, J. Hua, L. Luo, and S. Li, "Elnet: Automatic classification and segmentation for esophageal lesions using convolutional neural network," *Medical Image Analysis*, vol. 67, p. 101838, 2021.

[16] R. Ge, G. Yang, Y. Chen, L. Luo, C. Feng, H. Zhang, and S. Li, "Pv-lvnet: Direct left ventricle multitype indices estimation from 2d echocardiograms of paired apical views with deep neural networks," *Medical image analysis*, vol. 58, p. 101554, 2019.

[17] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn *et al.*, "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: a cancer journal for clinicians*, vol. 69, no. 2, pp. 127–157, 2019.

[18] Y. Wang, X. Lu, Y. Zhang, X. Zhang, K. Wang, J. Liu, X. Li, R. Hu, X. Meng, S. Dou *et al.*, "Precise pulmonary scanning and reducing medical radiation exposure by developing a clinically applicable intelligent ct system: Toward improving patient care," *EBioMedicine*, vol. 54, p. 102724, 2020.

[19] "United imaging's emergency radiology departments support mobile cabin hospitals, facilitate 5g remote diagnosis." 2020. [Online]. Available: https://www.prnewswire.com/news-releases/united-imagings-emergency-radiology-departments-support-mobile-cabin-hospitals-facilitate-5g-remote-diagnosis-301010528.html

[20] R. Booij, M. van Straten, A. Wimmer, and R. P. Budde, "Automated patient positioning in ct using a 3d camera for body contour detection: accuracy in pediatric patients," *European Radiology*, vol. 31, no. 1, pp. 131–138, 2021.

[21] N. Saltybaeva, B. Schmidt, A. Wimmer, T. Flohr, and H. Alkadhi, "Precise and automatic patient positioning in computed tomography: avatar modeling of the patient surface using a 3-dimensional camera," *Investigative radiology*, vol. 53, no. 11, pp. 641–646, 2018.

[22] K. İncetan, R. Mohan, H. Stoutjesdijk, N. Fernandes, and B. de Jager, "Rgb-d camera-based clinical workflow optimization for rotational angiography," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8867–8874, 2020.

[23] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7450–7459.

[24] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang, "Part-aware context network for human parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8971–8980.

[25] "Yolov5," *https://github.com/ultralytics/yolov5*.

[26] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[28] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[29] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[31] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[32] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[33] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[38] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *In Proc. CVPR*, 2013.

[39] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[40] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[41] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[42] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6982–6991.

[43] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1831–1840.

[44] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6449–6458.

[45] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103–7112.

[46] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[47] B. M. H. Romeny, *Front-end vision and multi-scale image analysis: multi-scale computer vision theory and applications, written in mathematica*. Springer Science & Business Media, 2008, vol. 27.

[48] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.