Video2Report: A Video Database for Automatic Reporting of Medical Consultancy Sessions

Laura Schiphorst, Metehan Doyran, Sabine Molenaar, Albert Ali Salah, Sjaak Brinkkemper
Dept. Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

Abstract— **The regulation of medical consultations for some countries, such as the Netherlands, dictates the general practitioners to prepare a detailed report for each consultation, for accountability purposes. Automatic report generation during medical consultations can simplify this time-consuming procedure. Action recognition for automatic reporting of medical actions is not a well-researched area, and there are no publicly available medical video databases. We present in this paper Video2Report, the first publicly available medical consultancy video database involving interactions between a general practitioner and one patient. After reviewing the standard medical procedures for general practitioners, we select the most important actions to record, and have an actual medical professional perform the actions and train further actors to create a resource. The actions, as well as the area of investigation during the actions are annotated separately. In this paper, we describe the collection setup, provide several action recognition baselines with OpenPose feature extraction, and make the database, evaluation protocol and all annotations publicly available. The database contains 192 sessions recorded with up to three cameras, with 332 single action clips and 119 multiple action sequences. While the dataset size is too small for end-to-end deep learning, we believe it will be useful for developing approaches to investigate doctor-patient interactions and for medical action recognition.**[1]

I. INTRODUCTION

In the Dutch healthcare system, care providers (CPs) are obliged to accurately report on the encounters with their patients and on the treatments in an electronic medical record (EMR). These EMRs are designed for improved communication between CPs and capture previous diseases, treatments, and observations [8], [22]. Moreover, they serve to comply with guidelines and can support medical decisions [2]. Even though the EMRs support the medical care for patients, accurately documenting all aspects of healthcare is time consuming, since it is done manually by the CPs. Administration tasks in healthcare are estimated to take over 100,000 full-time positions in long-term care in the Netherlands with a total cost exceeding 5 billion euros per year[2]. A more efficient and less time-consuming way of reporting medical consultations is necessary. Automatically constructing and storing medical reports in the EMR may be a solution. Recognising actions from videos could aid in automatically constructing these reports and recent developments in action recognition provide promising results in other fields.

This study aims for recognising actions from videos of medical consultations. To do so, a suitable dataset on medical actions is required. However, to our knowledge, no datasets consisting of one-on-one interactions between GPs and their patients are publicly available. Therefore, in this work, we design and collect Video2Report, a database of medical actions, with conditions similar to real consultation scenarios.

This work is part of the Care2Report Project[3] [16] that aims to automatically report and document the medical documents in the EMR by combining automatic speech recognition and action recognition to recognise the relevant medical actions that are performed during consultancy sessions.

We focus here primarily on human-human interactions between GPs and their patients, rather than a team of specialists operating simultaneously. More precisely, we aim to recognise the set of most frequently performed medical actions performed during a medical consultation, such as blood pressure measurement and auscultation of the heart and lungs. The automatic action recognition results can then be converted into a text-based report draft, which will be completed (and corrected) by the practitioner to eventually add all relevant information to the EMR.

II. RELATED WORK

A. Databases

The action recognition literature saw rapid progress over the last years, with databases initially containing simple, single-person actions in isolated backgrounds [21], followed by increased variance over recording conditions and including simple two-person actions, like fighting and meeting [6], [1] or a combination of single- and multiple person actions [18], [13], [17]. As described in a recent survey, the number of classes have increased to hundreds of actions in the larger action recognition databases, and the number of clips can exceed a million [26]. However, these large sets are typically harvested from large multimedia websites such as YouTube, where the recording conditions cannot be easily controlled, the actions are not scripted, and the annotations are costly to create.

For the existing interaction datasets, it is rare to have multiple viewpoints at the same time (many are harvested from movies and thus have single and often moving camera), and

This work was supported by the Care2Report project.

[1]This is the uncorrected author proof. Copyright with IEEE, please cite as: Schiphorst, L., M. Doyran, A.A. Salah, S. Molenaar, S. Brinkkemper, "Video2Report: A Video Database for Automatic Reporting of Medical Consultancy Sessions," 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, 2020.

[2]https://www.berenschot.nl/actueel/2016/juli/administratieve-taken/

[3]www.care2report.nl

while efforts like the Panoptic Studio at CMU [12] provide large amounts of data by recording simultaneously from tens of cameras, they are expensive to create, and do not yet contain typical actions of a medical practitioner. Similarly, several datasets exist with multiple camera recordings in the surveillance domain, where scenes with camera angles suitable for CCTV cameras are used [24]. A good overview with image samples is provided in [4]. A video database with scripted medical interactions is currently lacking.

B. Approaches

State of the art approaches for action recognition in dyadic human-human interactions are based on convolutional neural networks (CNNs) [26]. A very popular approach is the two-way CNN, where the camera image and an optic flow image are fed in parallel to a CNN [23]. However, even in restricted setups, the training of such architecture require at least on the order of 10K videos.

For scenarios where much less training data are available, the most straightforward approach is to use a CNN trained on a different database as a feature extractor, and complement it with a classifier specifically trained for the task [5]. Long short-term memory (LSTMs) networks are typically used to model temporal dynamics [7], [11], [15], [27].

Skeleton data are one of the most commonly used features for action recognition. In [19], the authors propose an approach for using joint angles from three dimensional (3D) skeleton features to recognise human actions with a linear support vector machine (SVM). Spatial features and spatio-temporal features are extracted from the 3D skeleton joints. Song et al. [25] trained an LSTM network to recognise which joints were dominant in certain actions, using a temporal attention module. Zhang et al. [28] extracted geometric relations amongst all joints from the 3D skeletons. We will also use skeleton features in our baseline approach.

III. VIDEO2REPORT DATABASE

To collect the Video2Report medical action database, we looked at the literature to select the most relevant actions for the general practitioners, and enlisted the help of a medical expert for the recordings, who further trained two novices in the proper procedures. This ensured that the actions were performed with natural movements. Consequently, we had three different persons performing the actions. We used the publicly available ELAN video annotation tool, to annotate the data [10]. Each session was annotated by a single annotator, as the procedures are scripted and clear.

A. Selection of actions

The sessions consist of one-on-one encounters between GPs and their patients. In order to best represent a real consultation, we use existing clinical guidelines for Dutch health practitioners. These are available online at the website of the Dutch Health Practitioners Society (Nederlandse Huisartsen Genootschap)[4].

[4]https://www.nhg.org/nhg-standaarden

TABLE I: Medical action statistics in our dataset.

Medical action	# of clips	Mean Length (sec.)
Blood pressure measurement	124	75.6
Palpation abdomen	94	22.7
Percussion abdomen	84	12.5
Auscultation lungs	155	32.7
Auscultation heart	145	24.8
Auscultation abdomen	96	17.1
No action	451	6.19

From these guidelines, we found the most common medical actions and treatments for ninety one syndromes. We eliminated the medical actions that are either considered too private to record, that consisted of inspection with the eyes by the GP, or for which we need medical equipment that we did not have at our disposal.

A MySignals kit [14] was used for blood pressure monitoring, and a red and a black stethoscope were used for auscultation purposes. When providing the baseline results, we use geometrical features to ensure that visual features of such equipment are not memorized by the classifiers.

Each session represents one medical consultation, and may contain a single action or multiple actions in a sequence. Medical actions that are combined most often during a consultation are auscultation of the heart and lungs; as well as auscultation, percussion, and palpation of the abdomen. In an orienting physical examination, these five medical actions are combined in a single consultation. Measuring blood pressure is often performed together with auscultation of the heart or the lungs. Consequently, these are selected to be the medical action classes. A 'no action' class was added to these.

The dataset composition is detailed in Table I, which shows the number of clips and mean length per action in the database. Note that some actions co-occur in some of the clips, as a sequence of medical actions. These medical actions occur in 42 of the medical procedures, accounting for a total of 46% of all the medical guidelines we investigated. Informed consent is obtained on all recordings. No actual patients participated in the data collection. The database and all annotations are made publicly available[5]. All the faces in the videos will be blurred to preserve the privacy of the medical professional.

B. Recording of the sessions

To be able to learn actions from different viewing angles and to model different recording conditions, we use multiple cameras while filming the actions. Additionally, we can also investigate what position of the camera is most convenient to use for a practical setup. We used a Panasonic HC-V770 (which is referred to as the 'camera' from now on), a GoPro Hero 5, and an iPad. The videos have a resolution of 1920×1080 pixels, with a frame rate of 30 fps.

We decided to position the cameras at different heights and in different locations. In order to create the maximum overview with the least amount of occlusion, the camera is

[5]https://github.com/dmetehan/Video2Report

positioned slightly higher, such that a bird's eye view is created. The GoPro has a 170° angle and is positioned at eye height. The iPad is positioned in the corner of the room, at eye height (Fig. 2). Fig. 1 represents images at the same moment in time, captured by the three different cameras.

Variations in the medical actions include positioning of the GP with respect to the patient, palpation order (left to right or right to left), GP's movements (clockwise or counterclockwise), and different (random) order in addressing the area of the abdomen. Moreover, during auscultation of the lungs, the GP can start by either listening to the right or to the left side of the patient's body. Patients also showed variations in poses, with bent legs or stretched out during the examination. Four different subjects (3 females and 1 male with average age of 26.25) participated in the recordings, which spanned multiple sessions on different days, and the subjects have individual variations in clothes, hair styles, and accessories. Note that GPs in the Netherlands wear regular clothes, rather than a doctor's coat, so it is not possible to distinguish them by clothing (except when they are wielding a stethoscope).

In Fig. 1 we show three different views for one of the sessions. Even though the images are from the exact same moment, there are differences in lighting conditions, camera angle, and zooming options. For example, the recordings of the camera has a warmer lighting compared to the GoPro, which also depends on the camera's optics and correction. The GoPro has a wide-angle lens and is more zoomed out, compared to camera and iPad, creating an overview of the entire setting. In any given condition, there will be occlusions in the videos, as the GP may walk around the patient, but most important actions will take place within the viewing angle of the cameras.

C. Annotations

We annotated the videos manually, using the publicly available video annotation tool ELAN [10]. We annotated 1) the posture of the patient, 2) the distance from the GP to the patient (i.e. touching vs. not touching), 3) the area of investigation, and 4) the medical action.

The rules observed during the annotations are as follows:

1) The medical action starts from the moment the GP touches the patient, either with the hands or with a medical instrument. It lasts until the GP no longer touches the patient.
2) The GP is considered to touch the patient either when the hands or a medical instrument touches the patient at the part of the body where examination takes place.
3) The area of investigation is the part of the patient's body where the medical action takes place, and it is annotated for the entire duration of the medical action. An exception is made for blood pressure measurement, for which we annotate the area of investigation for the duration of the medical action, as well as for only when the GP is considered to touch the arm. We have annotated this as 'Arm' and 'ArmTouch', respectively.
4) The posture of the patient is only defined at the static moments, not in the transition phase. 'Sitting upright' is annotated when the patient body is vertical, while 'lying down' is annotated when the patient's body is fully horizontal.

D. Database Content

In total, we recorded 192 unique sessions: 28 sessions with one, 69 with two and 95 with three cameras simultaneously, accounting for a total of 451 videos. Out of these videos, 332 contain a single action, while the remaining 119 consist of sequences of actions.

The majority of the videos were recorded with both a female GP and patient (131 sessions, 68.2%), while 15.6% of the videos have a female-GP/male-patient (30 sessions), and 16.7% of the sessions have a male-GP/female-patient distribution (31 sessions).

E. Privacy

Privacy is of great importance for the project. While actions like undressing mostly happen behind a closed curtain, the intimacy of some medical actions may appear inappropriate for the patient to film. Moreover, patients may feel uncomfortable being recorded during their consultation in general, since they may be discussing private issues. In the project, the planned system only stores the videos shortly, until the classification results are generated, which should be near real-time. Patients should clearly see that they are being recorded, should see -if they wish- what is being recorded, and should be informed that the recordings are not stored beyond the session, as this may influence their sense of feeling safe and secure.

IV. BASELINE EXPERIMENTS

We extracted skeleton features from the videos using the widely popular OpenPose system [3], which provides 25 body landmarks for each person in each frame (Fig. 3). We apply nearest neighbor matching to track the skeletons and smooth the locations using a Savitzky–Golay filter [20].The raw body landmark locations are processed to calculate distances and angles for each skeleton, and more importantly, distances between the doctor's hands and the patient's landmarks. The latter provides valuable information on the area of investigation, which is crucial for the correct classification of medical actions. These static features are then pooled using functionals such as mean and variance over a 30-frame sliding windows with 15 frame skip. This provides temporal features in the form of the amount of change in distance and angle between landmark locations.

We have contrasted four classifiers for the baseline experiments: Decision Trees (DT), Random Forests (RF), Extreme Learning Machines (ELM), and Long-Short Term Memory (LSTM) networks, respectively. While we did not train end-to-end deep neural networks, our feature extraction (i.e. OpenPose) uses a two-way convolutional neural network.

In the DT classifier, we have used Gini impurity measure [9] for branching. We have used categorical cross-entropy loss in the LSTM network, and for ELM, a plain

Fig. 1: Images from the same session at the same moment in time, captured by the three different cameras. Left: camera, middle: GoPro, right: iPad

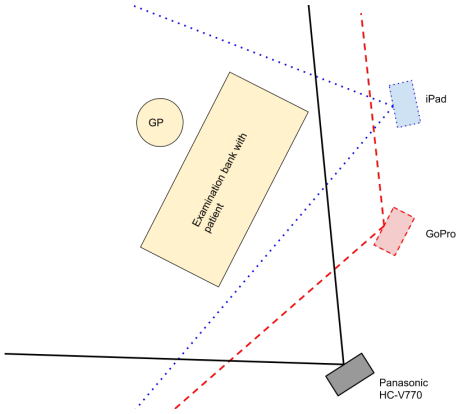

Fig. 2: Field of view for the setup of the recording sites.

Fig. 3: Illustrating skeletons extracted via OpenPose.

Fig. 4: Confusion matrix of the best performing method

TABLE II: Baseline results (accuracy)

Method	DT	RF	ELM	LSTM	Majority
Mean	61.82	69.13	65.35	68.67	33.66
Standard deviation	2.15	2.08	3.65	2.96	3.25

version was used, where the initial layer weights are randomly selected. The rest of the parameters for the classifiers are described in the next section.

V. EXPERIMENTS AND RESULTS

We split the dataset into four folds. Three parts are used as training (including validation), and one part is used for testing. Parameter search is conducted within the training fold, by a 4-fold cross-validation. We report mean accuracy (and standard deviation) for action recognition with the best parameters for each method (as learned within the training fold) in Table II. As a final baseline, we add the Majority predictor to the table, which predicts only the most occurring action class, which is blood pressure measurement.

The best parameters found by the grid search (on the first training fold) are as follows: DT with maximum depth of 50, maximum leaf nodes of 250; RF with 25 trees all with maximum depth of 50; ELM with 10K hidden units and tanh activation function; LSTM with 250 hidden units plus a fully connected layer with 100 hidden nodes.

In Fig. 4 we visualise the confusion matrix of the best performing baseline method to inspect the further details of the classification problems. The highest confusion can be seen between the actions auscultation of the abdomen and palpation of the abdomen. Similarly, auscultation of the heart and lungs, as well as between percussion and palpation of the abdomen are confused. These are similar actions, performed in the same area of investigation. Other then area of investigation related problems, blood pressure measurement is occasionally confused with the 'no action' class. All the other actions have clear movements, but blood pressure measurement has very little movement in it. Doctors also tell the patients not to move during this measurement. It may be that using only skeleton-based features are not enough to capture this action, and additional image-based features should be used.

VI. CONCLUSIONS

We collected and annotated the first publicly available medical consultancy video database. Medical actions have lower variability, smaller activity space and less movement than the most of the actions that we see in other domains, such as sports or daily actions. The reported baseline results illustrate the need for a more precise area of investigation localisation.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of B.Sc. M. N. Van Lingen for the dataset collection.

REFERENCES

[1] S. Blunsden and R. Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4, 2010.

[2] P. Campanella, E. Lovato, C. Marone, L. Fallacara, A. Mancuso, W. Ricciardi, and M. L. Specchia. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. *The European Journal of Public Health*, 26(1):60–64, 2015.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[4] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[6] R. B. Fisher. The pets04 surveillance ground-truth data sets. In *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–5, 2004.

[7] H. Ge, Z. Yan, W. Yu, and L. Sun. An attention mechanism based convolutional lstm network for video action recognition. *Multimedia Tools and Applications*, 78(14):20533–20556, Jul 2019.

[8] J. F. Golob Jr, J. J. Como, and J. A. Claridge. The painful truth: The documentation burden of a trauma surgeon. *Journal of Trauma and Acute Care Surgery*, 80(5):742–747, 2016.

[9] J. L. Grabmeier and L. A. Lambe. Decision trees for binary classification variables grow equally with the gini impurity measure and pearson's chi-square test. *International Journal of Business Intelligence and Data Mining*, 2(2):213–226, 2007.

[10] B. Hellwig. *ELAN - Linguistic Annotator*. The Language Archive, MPI for Psycholinguistics, Nijmegen, The Netherlands, version 5.4 edition, 12 2018.

[11] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. pages 1971–1980, 06 2016.

[12] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[13] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. 2008.

[14] Libelium Comunicaciones Distribuidas S.L., Zaragoza, Spain. *MySignals SW eHealth and Medical IoT Development Platform Technical Guide*, 4.6 edition, 5 2019.

[15] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[16] L. Maas, M. Geurtsen, F. Nouwt, S. Schouten, R. Van De Water, S. Van Dulmen, F. Dalpiaz, K. Van Deemter, and S. Brinkkemper. The care2report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[17] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*, pages 2929–2936. IEEE Computer Society, 2009.

[18] A.-T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 476–481. IEEE, 2007.

[19] E. Ohn-Bar and M. Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 465–470, 2013.

[20] W. H. Press and S. A. Teukolsky. Savitzky-golay smoothing filters. *Computers in Physics*, 4(6):669–672, 1990.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[22] J. E. Sheppard, L. C. Weidner, S. Zakai, S. Fountain-Polley, and J. Williams. Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. *Archives of disease in childhood*, 93(3):204–206, 2008.

[23] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[24] S. Singh, S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 48–55. IEEE, 2010.

[25] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[26] A. Stergiou and R. Poppe. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding*, 188:102799, 2019.

[27] M. Wang, B. Ni, and X. Yang. Recurrent modeling of interaction context for collective activity recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7408–7416, July 2017.

[28] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157. IEEE, 2017.