Fine segmentation for Activity of Daily Living analysis in a wide-angle multi-camera set-up

Philipe Ambrozio Dias[1]
philipe.ambroziodias@marquette.edu
Henry Medeiros[1]
henry.medeiros@marquette.edu
Francesca Odone[2]
francesca.odone@unige.it

[1] Department of Electrical and Computer Engineering
Marquette University
Milwaukee, WI, USA

[2] DIBRIS
Università degli Studi di Genova
Genova, Italy

Abstract

This paper presents a fine segmentation pipeline, designed as a building block of an Activity of Daily Living analysis framework. The reference application domain is a protected discharge residence, where elderly patients spend a few days and their health status and frailty condition is assessed continuously and automatically. The method we propose addresses a variety of challenges which are typical of the application we consider. It provides accurate segmentation, which will help the estimation of human-object and human-human interaction. It is tolerant to occlusions and geometric deformations, and it can detect objects of interest that are not in the foreground of the scene. We report promising quantitative results both on the benchmark DAVIS dataset and on video-streams acquired in the facility.

1 Introduction

With the rapid population aging occurring worldwide, there is increasing interest in estimating the health status and frailty of the elderly. *Frailty* is a condition of increased risk of negative health outcomes, including institutionalization, hospitalization and death, due to multi-systemic impairments in multiple domains such as physical, cognitive, and social [1]. In order to identify, measure and monitor frailty, geriatricians adopt various Prognostic and Frailty Indices and use them on hospitalized and community-dwelling older people when performing regular check-ups. Such indices combine information on the clinical, cognitive, functional, nutritional, and social skills and are collected through questionnaires as well as clinical exams and performance tests [2, 3, 4, 5].

No technology currently exists that allows geriatrically-relevant parameters of elderly patients to be monitored unobtrusively over long periods of time in a long-term care facility or at the patient home. In order to fill this technological gap, together with physicians from the Galliera Hospital (Genoa, Italy), we have designed a protected hospital-discharge facility, equipped as a comfortable two-bedroom apartment in which different sensors including

video-cameras have been arranged and properly hidden in order to enable an unobtrusive monitoring of the subjects who spend a few days in the facility after discharge from the hospital. As a core requirement to achieve unobtrusive monitoring, the long-term goal of our work is to create video analytics tools to robustly and accurately measure relevant mobility parameters of elderly patients in a relatively uncontrolled environment.

In this paper we address a fundamental building block of this long-term research: the precise segmentation and tracking of individuals in video-streams acquired by a multi-camera system. Fine-grained segmentation is crucial for the subsequent steps of the procedure, namely action and activity recognition needed to address Activity of Daily Living (ADL) analysis. In order to gather complementary information on the surrounding environment, to be used to improve the quality of our analysis, we also identify objects belonging to a set of pre-defined classes of interest. Currently we are considering the following objects: sofa, chair, and dining table. Our precise segmentation will allow us to obtain robust models of person-person and object-person interaction to be used within the ADL analysis, to access functional abilities (that is, the ability of using tools), independence, and social awareness.

The main challenges of our work are the complexity of the scenario and the fact that the objects of interest may appear in different parts of the image, at different scales, poses and deformations — because of significant distortions due to the need of adopting wide angle optics (see Figure 1). To address these challenges, we propose a multi-stream network in which different patches of a video-frame are fed to separate copies of the network. View-specific distortions are taken into consideration by applying simple ad-hoc geometric transformations to the image patches. The outputs of the different branches of the network are combined using a fusion mechanism and refined using superpixel segmentation and a probabilistic temporal consistency model.

Figure 1: Example of frame containing perspective-distortion. By applying view-specific transformations (e.g. 45° rotation) to the distorted regions, the segmentation can be substantially improved.

We assess the performance of our method on the benchmark DAVIS (Densely Annotated VIdeo Segmentation) dataset [6, 7] as well as on video streams acquired within our protected discharge facility. The results show that our approach outperforms state-of-the-art video segmentation methods on selected sequences of the standardized datasets and that it generates very accurate semantic object segmentation in the real-world videos.

2 Related work

In order to make inferences about activities of daily living, accurate knowledge about the spatio-temporal relationships among people and objects is needed. Therefore, the ability to classify each image pixel, a task known as semantic image segmentation, is an important element of smart environments. When such segmentation must be carried out in a temporally consistent manner across video frames, as is the case in such scenarios, this task is known as semantic video segmentation.

Semantic segmentation of images has been subject of computer vision research for many years, but the introduction of publicly available datasets such as MSRC-21 [8] and the PASCAL VOC challenge [9] accelerated research in this area. Similarly, semantic video segmentation has also benefited from the introduction of publicly available datasets such as DAVIS [6, 7] and SegTrack [10]. Most recent works on video segmentation, however, focus on the simpler task of object segmentation, which consists of accurately segmenting an object in every frame of a video sequence given a mask representing this object in the first video frame. In this section we briefly survey some of the most relevant works in semantic image segmentation as well as in object and semantic video segmentation.

As in most areas of computer vision, earlier semantic image segmentation methods employed features based on color and texture in conjunction with bag-of-visual words classifiers [11] or conditional random fields [12], but most recent methods are based on features extracted by convolutional neural networks [13, 14, 15]. One particularly successful recent strategy is based on the application of fully convolutional networks (FCNs) [16, 17, 18, 19, 20]. In these approaches, rather than terminating the network using fully connected layers followed by a softmax layer to perform classification, the outputs of the convolutional layers are upsampled to generate a dense spatial map of pixel labels.

Different upsampling approaches have been employed in order to preserve fine segmentation details, including the application of deconvolution layers [20, 21], encoder-decoder architectures [16, 18], and the use of dilated (or *atrous*) convolutions [17]. Although the 'atrous' strategy generates accurate results by capturing higher level information using larger receptive fields rather than downscaling the image, it is computationally expensive.

Significantly less computationally expensive methods have shown similar performances according to the PASCAL VOC 2012 segmentation challenge [1]. RefineNet [19] is one such method. As most of the best performing recent approaches, it leverages the ability of residual networks [22] to learn deeper and more complex representations of images. RefineNet [19] employs a multi-path refinement structure such that long-range residual connections are formed. Each block has as input two of the feature maps collected from the residual network at resolutions of 1/4, 1/8, 1/16 and 1/32 of that of the original image. The inputs are combined through a sequence of adaptive convolutions for task-specific fine tuning, upsampling for multi-resolution fusion and residual pooling to capture background context.

Although most of the methods described above show impressive performances on publicly available benchmark datasets, their main limitation is the fact that they focus on the segmentation of a few (frequently only one or two) salient and large foreground objects. Although the more recent PASCAL-context dataset [23] introduces more complex scenarios involving more objects, the performance of most existing approaches on this new dataset is still far from satisfactory. Addressing this limitation is one of the objectives of the proposed

[1] The PASCAL VOC dataset is the most widely used dataset for visual object segmentation, which makes its leader board a good reference to quantitatively compare existing state-of-the-art methods.

work.

Regarding video segmentation, although some recent approaches which are not based on CNNs have shown good performance [24], methods that employ CNN features to model the appearance of the object tend to perform better [25]. Again, approaches based on FCNs also dominate this field [26, 27, 28]. One-shot video object segmentation (OSVOS) [26], for example, uses the FCN of [21] to carry out object segmentation in a frame-by-frame basis without imposing temporal constraints. MSK (MaskTrack+Flow+CRF) [28] performs object segmentation using the DeepLab network of [17] with the object segmentation masks of the previous frame provided as a fourth input channel to the network in order to take into consideration the temporal information. In [27], Jampani et al. propose the Video Propagation Network (VPN), which is one of the few recent approaches to perform both semantic and object video segmentation. Their method uses a bilateral filtering network [29] to carry out temporal propagation and a CNN for spatial segmentation. It has also been integrated with other FCNs such as DeepLab [17].

In addition to the fact that most of the above methods solve the problem of object segmentation in videos, disregarding semantic information, as with the semantic image segmentation methods, they focus on segmenting a few large and prominent foreground objects from the background, and hence cannot be directly applied to monitoring the scenarios under consideration.

3 Our approach

The method we propose for semantic segmentation of video frames uses a core RefineNet model to compute the likelihood that each pixel belongs to a certain category of interest. In particular for this preliminary study, we opted for the ResNet-101 based model that provides state-of-the-art performance on the PASCAL VOC 2012 dataset, which includes all the objects classes we currently consider for ADL analysis: *person, chair, sofa, dining table*.

Since our video-sequences for ADL analysis consist of frames acquired with static wide-angle cameras, we also incorporate in our approach strategies to compensate for distortions and to detect non-centered objects. Specifically, instead of evaluating only the entire input frame, we devise a semantic video segmentation architecture based on a spatial multi-stream arrangement (Fig. 2). In each stream, a region is cropped from the input frame and fed into a RefineNet module, which outputs 20 pixel-dense feature maps corresponding to the likelihood that a pixel belongs to a certain PASCAL class. For each class, the computed scores are then combined by a late fusion layer. In addition, in order to compensate for perspective distortions from our wide-angle cameras, each portrait is evaluated both with and without a 45° counter-clockwise rotation. After adding the responses obtained from all the portraits, a pixelwise maximum likelihood evaluation indicates to which class a pixel most likely corresponds.

Conceptually, this type of architecture allows an adaptive feature extraction arrangement in which each CNN module composing a stream as well as the late fusion layer can be fine-tuned through task-specific training. We envisage view-specific training as an important part of our future work. In our current proof-of-concept implementation no supervised fine-tuning is performed, such that all the RefineNet modules share the same pre-trained weights, view-specific segmentation is approximated by the 45° rotations, and late fusion is performed by a simple summation.

Another important requirement in our reference application is the accuracy of the seg-

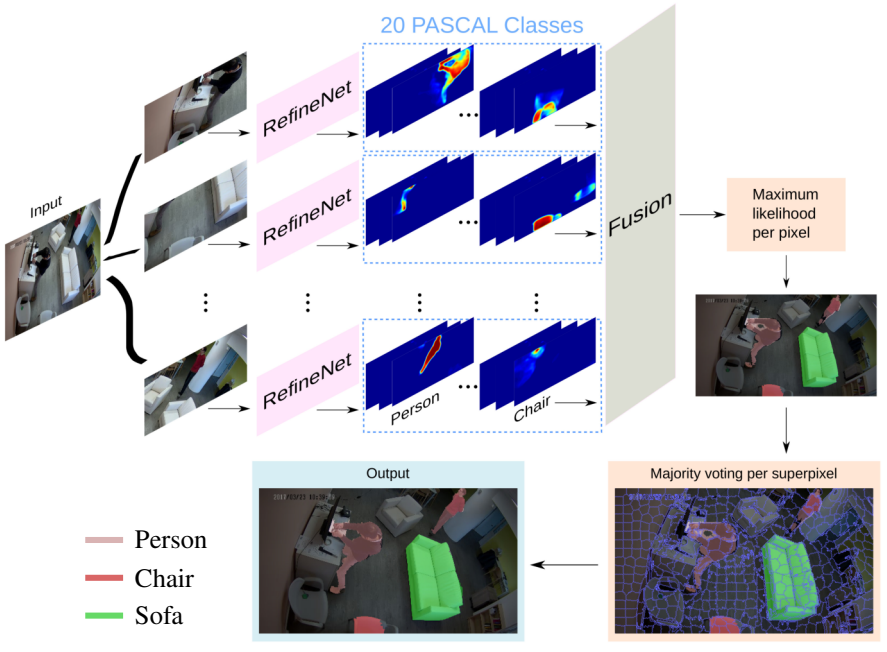

Figure 2: Diagram illustrating the sequence of image analysis performed by the proposed model for semantic segmentation of objects of interest.

mentation so that interactions between different objects and agents can be reliably estimated. Although the segmentations obtained with RefineNet are finer than the ones obtained with deconvolutional models such as FCN, a close inspection reveals that the results can be improved especially in terms of boundary adherence. Unsupervised techniques such as superpixel segmentation are capable of better exploring local information to estimate boundaries of objects composing an input image. Therefore, in our approach, we additionally segment the images using superpixels and each superpixel is then classified according to a majority voting scheme based on the scores obtained from the RefineNet method. That is, if more than 50% of the pixels composing a superpixel present a score over a certain threshold for a given class, the superpixel is considered a positive detection of the object represented by that class.

In [30], Stutz et al. provide a review of existing superpixel approaches, with a comprehensive evaluation that ranked 28 state-of-the-art algorithms according to several metrics such as average recall, average undersegmentation error, boundary recall and also realtime capability. The energy optimization based algorithm Extended Topology Preserving Segmentation (ETPS) [31] stands out in their evaluation with high performance in terms of boundary adherence, recall, stability and runtime. For this reason, we selected this method to compose our image segmentation model.

For images acquired from our discharge facility, in addition to framewise analysis temporal information can be used based on prior information estimated from the environment. Since the images are acquired by static cameras, some objects of interest such as pieces of furniture have lower probabilities of moving between frames. Therefore, for these video

sequences we also include a simple probabilistic model that, in addition to the likelihood scores obtained using RefineNet, takes into account temporal information by attributing a higher probability of detection to pixels detected in the previous frame, while the probability of transition between labels (e.g. background to foreground or vice-versa) is lower.

4 Assessment on benchmark data

Quantitative evaluation of video segmentation methods requires pixel-accurate and per-frame ground truth annotation, a notoriously labor-intensive and time-consuming task. Given these difficulties, our application specific dataset has not been fully labeled by the time of this publication. For that reason, we quantitatively assess the performance of our method on video sequences composing the DAVIS 2016 dataset, which reflects many of the properties of our reference application. It comprises scenarios such as target occlusion, motion-blur, scenes with depth and appearance/pose changes, all of which are likely to occur in our application-specific video sequences. For an evaluation that resembles the environment of our application, where the classes of objects to be detected are known a priori, we selected only video sequences where targets correspond to objects contained on the PASCAL VOC 2012, disregarding sequences containing unknown objects.

To verify the efficacy of the proposed per-superpixel majority voting scheme, we compare two approaches against the baseline methods: one composed only by RefineNet (which we refer to as RN) and one combining RefineNet and superpixel analysis (which we refer to as RS). Although we provide a comparison against multiple video segmentation techniques proposed for the DAVIS challenge, it is important to note that our goal is not a task-restricted model that aim to achieve top rank performance on this specific dataset. For this reason, unlike most reference methods, we do not perform any type of fine-tuning using the training sequences provided by the DAVIS dataset. The only additional information used as prior knowledge are the classes composing the foreground/target of each sequence.

Following the official guidelines for the DAVIS Challenge, we compare our method against the baseline ones in terms of *Jaccard index* ($\mathcal{J}$) and *contour accuracy* ($\mathcal{F}$). The first is defined as the *intersection-over-union (IoU)* of an output segmentation and the ground-truth mask. The contour accuracy is defined as $\mathcal{F} = \frac{2P_cR_c}{P_c+R_c}$, where P_c and R_c stand for contour-based precision and recall, respectively. Table 1 summarize the results obtained for each video sequence, with the best results highlighted in bold[2].

For the evaluated sequences, both RN and RS approaches provide results that are competitive to the state-of-the-art methods, with average performance slightly superior for both metrics. This is particularly relevant considering that the baseline methods mostly have the advantage of being fine-tuned for this dataset. In addition, several of the sequences in the table were used as training sequences for some methods, and are hence not an indicative of their performance on data previously unseen by the trackers. For four video sequences, the performance in terms of segmentation similarity ($\mathcal{J}$) obtained using RefineNet based methods are superior to the ones provided by existing approaches. Similarly, for five sequences RN and RS achieve better contour accuracy ($\mathcal{F}$) than the baseline methods.

In addition, the performance of our method is consistent over time. Results obtained in terms of $\mathcal{J}$ decay ($\mathcal{D}_\mathcal{J}$) and $\mathcal{F}$ decay ($\mathcal{D}_\mathcal{F}$) evidence this characteristic as indicated in Figure 3 (Left), which shows that our method outperforms all the other approaches in these

[2]Note that results for OSVOS are not included in the table because the authors of [29] do not report their results on the official training set, which contains most of our selected sequences.

Table 1: Jaccard index (J) / Contour accuracy (F) Per-Sequence

Sequence	RS	RN	MSK [28]	VPN [27]	OFL [25]	BVS [24]	NLC [32]
bmx-bumps	45.8 / 60.5	42.6 / 63.0	57.1 / 67.8	41.8 / 59.2	47.5 / 52.9	43.4 / 49.3	**63.5** / **73.4**
bmx-trees	44.9 / 64.6	44.5 / 64.2	**57.5** / **73.6**	33.5 / 46.2	14.9 / 16.4	38.2 / 65.2	21.2 / 33.0
breakdance -flare	**86.4** / 91.2	83.5 / **92.2**	77.6 / 78.4	82.7 / 90.8	75.6 / 78.3	72.7 / 77.5	80.4 / 80.8
hike	90.6 / 94.2	85.5 / 94.8	93.1 / 96.0	88.0 / 95.4	**93.4** / **96.6**	75.5 / 76.4	91.8 / 94.3
hockey	83.2 / 81.4	80.8 / 83.1	83.4 / 79.1	78.5 / 80.3	**84.9** / **88.9**	82.9 / 85.0	81.0 / 80.8
horsejump -high	82.1 / 84.4	80.2 / 86.0	81.7 / 85.1	81.8 / 86.3	**86.3** / **90.4**	80.1 / 80.4	83.4 / 88.1
horsejump -low	82.6 / 86.8	**82.7** / **89.6**	80.6 / 81.2	74.4 / 71.3	82.2 / 85.9	60.1 / 56.5	65.1 / 65.9
kite-surf	64.7 / 44.8	60.7 / 42.5	60.0 / 43.8	62.3 / 53.5	**70.3** / 49.7	42.5 / **64.5**	45.3 / 44.8
lucia	89.9 / **92.4**	86.8 / **92.4**	**91.1** / 89.5	86.4 / 90.2	89.7 / 89.4	90.1 / 90.0	87.6 / 87.2
motocross -bumps	**89.2** / 81.9	88.0 / **82.8**	59.9 / 55.4	87.2 / 82.4	47.4 / 48.0	40.1 / 49.0	61.4 / 56.0
motorbike	79.1 / 75.6	79.9 / 75.3	56.6 / 59.7	**80.8** / **81.4**	47.6 / 50.4	56.3 / 69.6	71.4 / 57.1
paragliding -launch	61.4 / 20.6	59.2 / 19.4	62.1 / 22.9	61.4 / 23.1	63.7 / 25.3	**64.0** / **32.4**	62.8 / 24.3
parkour	**89.3** / 90.3	85.9 / **92.1**	88.2 / 87.4	87.3 / 91.7	85.9 / 87.0	75.6 / 67.8	90.1 / 91.6
rollerblade	84.5 / 86.1	81.7 / 89.5	78.7 / 85.0	81.4 / 87.9	**89.2** / **94.0**	58.8 / 64.5	81.4 / 86.8
scooter -gray	73.5 / 66.8	72.6 / **68.1**	**82.9** / 65.9	76.8 / 68.7	25.8 / 20.8	50.8 / 60.2	58.6 / 46.7
swing	77.0 / 66.2	75.1 / 66.3	81.9 / 74.5	82.5 / **78.7**	56.2 / 59.2	78.4 / 74.6	**85.1** / 77.8
tennis	85.1 / 90.5	82.5 / 92.1	86.1 / 91.1	79.0 / 89.4	81.7 / 87.2	73.7 / 84.5	**87.1** / **92.7**
MEAN	**77.0** / 75.2	74.8 / **76.1**	75.2 / 72.7	74.4 / 75.1	67.2 / 65.9	63.7 / 67.5	71.6 / 69.5

metrics. Moreover, a closer inspection based on pixelwise precision and recall (PR) metrics reveals that for most cases the detections provided by the proposed method are very precise. Figure 3 (Right) shows the PR curves summarizing the average performance of both RN and RS methods for the selected sequences. As the figure indicates, both approaches can simultaneously obtain precision and recall of approximately 85%, with the RS approach providing slightly higher precision at higher recall rates.

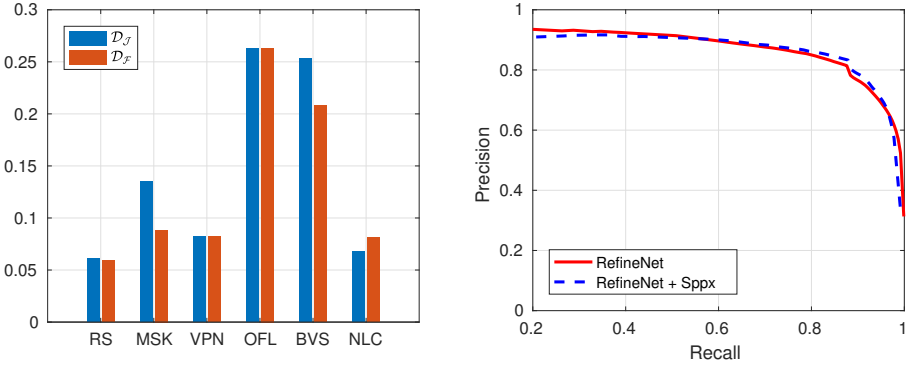

Figure 3: Performances on video sequences selected from the DAVIS 2016 dataset. *Left:* Mean $\mathcal{J}$ decay ($\mathcal{D}_\mathcal{J}$) and $\mathcal{F}$ decay ($\mathcal{D}_\mathcal{F}$) for each method. *Right:* Average precision recall curve of RefineNet based methods.

Figure 4 illustrates the segmentation accuracy for six scenarios that particularly resemble some challenges likely to occur in image analysis for ADL, e.g. poses variation, occlusion,

and scenes with depth and appearance changes. In these images, pixels correctly detected are marked in green. The red color indicates false positives, while false negatives are shown in blue. These results are a good evidence of the model robustness against such challenges. As illustrated by the frames extracted from the sequences *hockey* and *paragliding-launch*, false negatives mostly correspond to regions corresponding to objects unknown to the RefineNet model (i.e. not present in the PASCAL dataset).

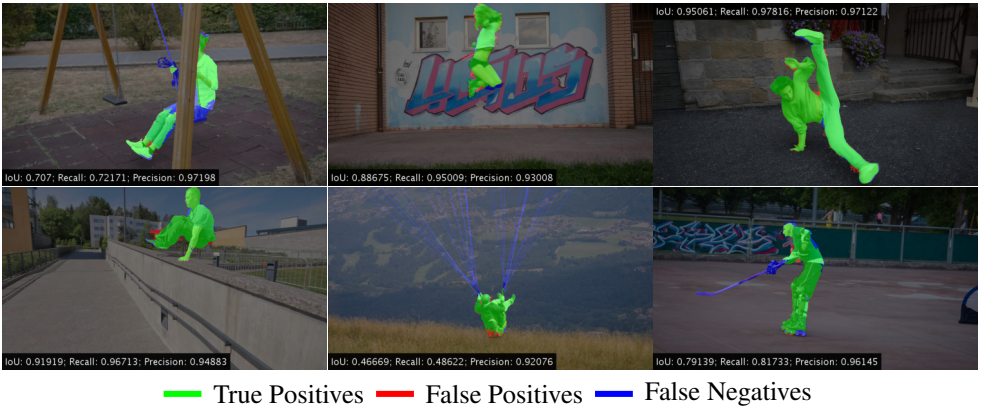

Figure 4: Examples of segmentation accuracy for scenarios including unusual poses, occlusion, depth and appearance changes.

5 Application to ADL

While the evaluation performed on benchmark data gives indications of the method performance in terms of the detection of objects of interest, the determination of its suitability for ADL analysis necessitates a task-oriented assessment using videos acquired within the protected discharge facility.

To quantitatively estimate the detection accuracy, we manually counted the number of correctly identified objects and false positives within two sequences of 50 frames, each acquired with one of the two cameras (named *view1* and *view2*) installed in our discharge facility. An object is considered correctly detected when at least 70% of its total area has been properly segmented, while a false positive corresponds to incorrect isolated detections of any size. To reduce labeling bias and in order to keep an approximately constant tolerance, each pair of detections was evaluated by the same human subject.

Two different approaches were evaluated. The first one (which we refer to as *RN*) consists of directly evaluating each frame using solely the RefineNet model, while the second (*MRST*) corresponds to the method we propose which is summarized in Figure 2 and employs multiple streams, superpixel enhancement, and the aforementioned temporal probabilistic model in conjunction with the pre-trained CNN.

Figure 5 shows qualitative results demonstrating that the proposed approach can detect and segment most of the relevant objects present within a scene, such as *person, chair, sofa, TV* and *table*. *MRST* in particular shows very promising results. Tables 2 and 3 summarize the quantitative results obtained for the video sequences acquired with cameras *view1* and

view2. Both tables present the total number of correct detections for each object category under consideration as well as the average number of correct detections per frame. The table also shows the total and average number of false positives. Both tables show a significant increase in the total number of correct detections for all object classes in both views. Although MRST generated 3 additional false positives in *view1*, all three occurred in the first three frames evaluated, before the temporal model stabilized.

Table 2: Analysis of 50 frames - View 1

	RN		MRST	
Class	Total	Avg.	Total	Avg.
People	136	2.72	141	2.82
Chairs	126	2.52	188	3.76
Tables	65	1.30	74	1.48
TV	19	0.38	38	0.76
FP	4	0.08	7	0.14

Table 3: Analysis of 50 frames - View 2

	RN		MRST	
Class	Total	Avg.	Total	Avg.
People	31	0.62	56	1.12
Chairs	48	0.96	50	1.00
Sofas	0	0.00	48	0.96
FP	9	0.18	0	0.00

Given the position of the second camera in the discharge facility (upper corner of the room), images acquired with this camera are particularly relevant since perspective distortions are present in every frame. The higher number of people and sofas detected in these frames using MRST demonstrate the effectiveness of the multiple streams and image rotations to obtain segmentations somewhat robust against the existing distortions.

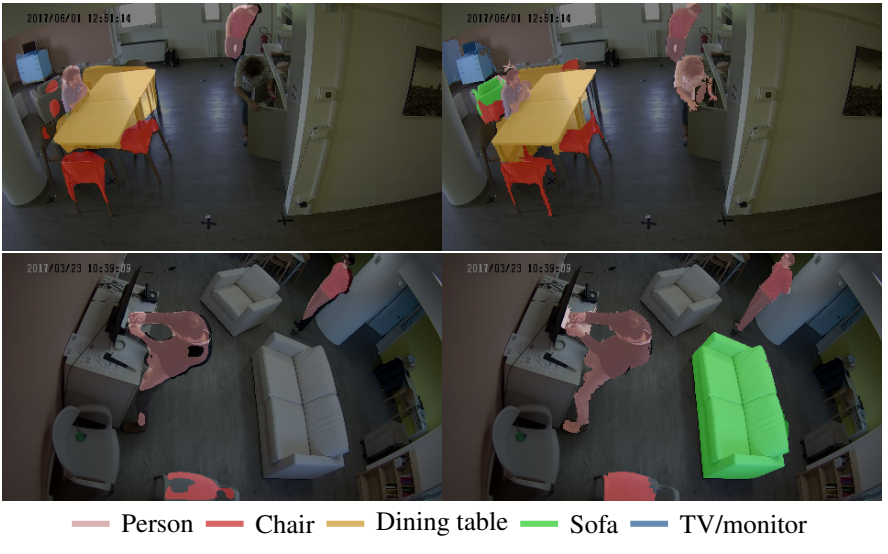

Figure 5: Examples of segmentation obtained for images acquired with cameras *view1* (top row) and *view2* (bottom row). *Left column:* results obtaining using solely RefineNet; *Right column:* results obtained combining multiple streams, RefineNet, superpixel enhancement and temporal probabilistic model.

6 Discussion and future work

We proposed a fine semantic video segmentation method for ADL analysis in multi-cameras assisted living applications. Our method employs the RefineNet semantic image segmentation approach in a multi-stream framework that allows the accurate segmentation of multiple objects in videos obtained using wide-angle cameras. Our approach further improves the segmentation accuracy by incorporating a superpixel majority voting post-processing mechanism as well as a temporal probabilistic model. Preliminary results show that our approach outperforms existing video segmentation methods in publicly available video sequences and performs accurate segmentation in a real-world assisted living facility.

In the future, we plan on extending our framework to fine-tune the different streams to perform view-specific segmentation so that ad-hoc transformations (e.g., 45° rotations) are not necessary. We also plan on incorporating more sophisticated temporal models that predict the expected position of objects in order to improve temporal consistency. More broadly, we plan on using the results of our semantic segmentation algorithms in activity recognition and tracking methods to generate automatic ADL analysis reports that can be used by geriatricians to evaluate the overall health status of their patients.

References

[1] L. P. Fried, L. Ferrucci, J. Darer, J. D. Williamson, and G. Anderson. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(3):M255–M263, 2004.

[2] N. M. De Vries, J. B. Staal, C. D. Van Ravensberg, J. S. M. Hobbelen, M. G. M. Olde Rikkert, and M. W. G. Nijhuis-Van der Sanden. Outcome instruments to measure frailty: a systematic review. *Ageing research reviews*, 10(1):104–114, 2011.

[3] P.A. Parmelee, P.D. Thuras, I.R. Katz, and M.P. Lawton. Validation of the cumulative illness rating scale in a geriatric residential population. *Journal of the American Geriatrics Society*, 43(2), 1995.

[4] A. Pilotto, L. Ferrucci, M. Franceschi, L. P. D'Ambrosio, C. Scarcelli, L. Cascavilla, F. Paris, G. Placentino, D. Seripa, B. Dallapiccola, et al. Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. *Rejuvenation research*, 11(1):151–161, 2008.

[5] M. E. Tinetti. Performance-oriented assessment of mobility problems in elderly patients. *Journal of the American Geriatrics Society*, 34(2):119–126, 1986.

[6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[7] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*, 2017.

[8] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[10] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[11] G. Csurka and F. Perronnin. A simple high performance approach to semantic segmentation. In *BMVC*, pages 1–10, 2008.

[12] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 739–746, Sept 2009. doi: 10.1109/ICCV.2009.5459248.

[13] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1915–1929, Aug 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.231.

[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. *Simultaneous Detection and Segmentation*, pages 297–312. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10584-0. doi: 10.1007/978-3-319-10584-0_20.

[15] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 82–90, Bejing, China, 22–24 Jun 2014. PMLR.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP (99):1–1, 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2699184.

[18] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[19] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017.

[20] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[21] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (4):640–651, April 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2572683.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[24] N. Maerki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral Space Video Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video Segmentation via Object Flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-Shot Video Object Segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] V. Jampani, R. Gadde, and P. V. Gehler. Video Propagation Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, july 2017.

[28] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung. Learning video object segmentation from static images. In *Computer Vision and Pattern Recognition*, 2017.

[29] V. Jampani, M. Kiefel, and P. V. Gehler. Learning sparse high dimensional filters: Image Filtering, Dense CRFs and Bilateral Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[30] D. Stutz. *Superpixel Segmentation: An Evaluation*, pages 555–562. Springer International Publishing, Cham, 2015.

[31] J. Yao, M. Boben, S. Fidler, and R. Urtasun. Real-time coarse-to-fine topologically preserving segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2947–2955, 2015.

[32] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. doi: http://dx.doi.org/10.5244/C.28.21.