A Fast Audiovisual Attention Model for Human Detection and Localization on a Companion Robot

Rémi Ratajczak, Denis Pellerin, Quentin Labourey
CNRS, GIPSA-Lab
Univ. Grenoble Alpes
F-38000 Grenoble, France
email: remi.ratajczak@gmail.com
email: denis.pellerin@gipsa-lab.grenoble-inp.fr
email: quentin.labourey@gipsa-lab.grenoble-inp.fr

Catherine Garbay
CNRS, LIG
Univ. Grenoble Alpes
F-38000 Grenoble, France
email: catherine.garbay@imag.fr

***Abstract*—This paper describes a fast audiovisual attention model applied to human detection and localization on a companion robot. Its originality lies in combining static and dynamic modalities over two analysis paths in order to guide the robot's gaze towards the most probable human beings' locations based on the concept of saliency. Visual, depth and audio data are acquired using a RGB-D camera and two horizontal microphones. Adapted state-of-the-art methods are used to extract relevant information and fuse them together via two dimensional gaussian representations. The obtained saliency map represents human positions as the most salient areas. Experiments have shown that the proposed model can provide a mean F-measure of 66 percent with a mean precision of 77 percent for human localization using bounding box areas on 10 manually annotated videos. The corresponding algorithm is able to process 70 frames per second on the robot.**

Keywords-audiovisual attention; saliency; RGB-D; human localization; companion robot.

I. INTRODUCTION

With the rapid advances in robotics, companion robots will tend to be more and more integrated in the human daily life [12]. These robots have the particularity to be both sociable, mobile and destined to evolve in an indoor domestic environment. One of the main requirement for them is to be able to quickly analyze their surrounding in order to interact with humans. That is why it is necessary to prioritize the perception, detection and localization of humans. They also need to behave as natural as possible in order to become acceptable presences for the humans [12].

To reach these requirements, cognitive based audiovisual attention mechanisms are a possibility that has been investigated in this work. Their related concepts can indeed provide the robot with the natural idea that it should give more attention to some positions than others.

A fast multimodal attention model for human detection and localization on a companion robot has thus been conceptualized and developed.

This model is distinctive from state-of-the-art methods presented in Section II due to both its application on a robot that can travel between different places during time, and to its architecture that combines visual, depth and audio data through two independent static and dynamic analysis paths. Moreover, since the robot and the humans can both move, the robot will not ever be in a situation where a specific characteristic of a human (face, leg, etc.) would be detectable. And since the detectors associated with these characteristics may sometimes fail, it has been decided not to use them in the proposed model in order to avoid false detections. This model may thus be considered as a bottom-up external information guided model.

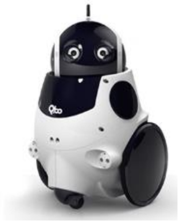

Figure 1. Photography of the robot Qbo Pro Evo without the Asus Xtion Pro Live RGB-D Camera.

It has been realized using the open source robot Qbo Pro Evo (Fig. 1) produced by TheCorpora©. Qbo's height is of 456 millimeters. It integrates an Intel i3-2120T 2.6 gigahertz processor and 4 gigabytes of random access memory. This hardware is conducted by a Linux Mint 17.1 operating system that has been enhanced with the Indigo's version of the Robotic Operating System (ROS). It embeds an Asus Xtion Pro Live RGB-D camera at the top of its head. This camera can stream depth and color images with a resolution of 640 by 480 pixels at 30 frames per second (FPS). It also provides two stereo microphones with a gap of 147.5 millimeters between them. Thanks to this system, the proposed model is able to analyze multimodal data as soon as they arrive.

The rest of this paper is organized as follows. Section II describes state-of-the-art ideas about audiovisual attention and its applications to robotics. Section III describes the proposed model in details. Section IV presents the dataset that has been used and the results of the evaluation. Section V concludes this study and presents future work perspectives related to the proposed model.

II. PREVIOUS WORK

This section presents previous work related to the proposed model.

A. *Audiovisual attention*

Audiovisual attention is a fast human cognitive process that aims to guide human interest through the most salient (i.e., remarkable) areas [7]. This process has been widely studied during the past three decades in neurosciences, psychology and computer sciences. This section focuses on the computational models developed using computer sciences methods.

Their goal is to represent the saliency level of different sources on a grayscale image called saliency map. On this map, the more salient an element is, the higher its intensity is [7]. Most attention models are referred as saliency models.

As described in [2], a huge number of saliency models have been implemented during time. Their efficiencies have been evaluated on different benchmarks using neurosciences ground truth results [3].

Anyway, these benchmarks are mostly available for the models designed for static two dimensional images only. They are not suitable to evaluate a multimodal or a dynamical model. This may be explained by the fact that most of state-of-the-art work are only based on static visual data (“single input image” [3]). These models have the particularity to give out salient regions independently of the content of the scene. This means that they may detect elements that do not correspond to a given target. In order to bias the results obtained with those classical models, some works have however incorporated other modalities, such as depth, motion, or sound.

Reference [5] demonstrated the utility to use a depth bias over two dimensional visual saliency results in order to increase their efficiency on ground truth evaluations obtained through eye-tracking processes. The authors notably concluded that humans are more attentive to close elements.

Reference [9] proposed to use motion detection on video images in order to localize areas that are moving and to combine them with a 2D static saliency model. The idea behind this is that the human gaze may be more attracted by moving objects than static objects. The authors proposed to drastically increase the saliency of moving objects.

Reference [4] has shown that adding sound analysis to visual cues may help to increase the saliency of a talking human for dynamic conversational purposes. Reference [10] used a visual additive two dimensional Gaussian bias centered on a horizontal sound localized position to improve the detection of a target in a complex visual environment.

B. *Applications in robotics*

Applications of the attention’s concepts in the field of robotics are still uncommon but are gaining more and more importance in the design of methods for robots’ perception.

Reference [11] proposed to combine a static visual attention model with a two dimensional sound localization to guide the gaze of a static humanoid robot thanks to the Head Related Transfer Function (HRTF) transform. This transform is effective to localize a sound in a human manner. However, it requires precisely designed humanoid ears covering the microphones, making the results of the HRTF difficult to reproduce with a standard robot such as Qbo.

Reference [12] used a multimodal approach to control the emotions of a sitting conversational humanoid robot according to the most interesting face of the human being. The authors used color, depth and sound data. In their method, they considered that the human faces will always be present in the scene. They used specific methods for human characterization such as emotion and head pose recognition. Their camera was detached of the robot and connected to an external computer. They did not consider algorithm speed issues.

III. PROPOSED MODEL

This section presents a novel approach for human detection and localization using audiovisual attention concepts on a companion robot. It has been inspired by the previously described independent literature results that have rarely been combined all together. The corresponding model thus combines multiple state-of-the-art ideas and methods in an all-in-one modular model represented on Fig. 2. It extracts independent static and dynamic features using visual, depth, and sound data. These features are then fused together in order to increase the saliency of the areas that may correspond the most to a human being.

In order to achieve this goal, the proposed model has been designed considering real domestic conditions through the following hypotheses: hypothesis (1) the robot will sometimes not be in presence of a human being, hypothesis (2) the robot may move over time, and will thus see different places with different points of view, hypothesis (3) a human being is a multimodal entity that can move and emit sound that the robot should be attentive to, hypothesis (4) the robot should be more attentive to close elements in order to avoid background salient elements detection, and hypothesis (5) the model has to be fast in order to eventually enable other processes to run at the same time on the robot. Its development was made considering the robot stationary while analyzing a scene. The mobility constraint has been considered through hypothesis (2).

A. *Processing architecture of the model*

As shown on Fig. 2, the proposed model has been decomposed in five steps and two independent static and dynamic analysis paths. It combines static visual 2D saliency with depth, motion and sound biases as referred in Section II.A. These modalities have been chosen according to the hypothesis (3) made in Section III. From a computational point of view, the model has been first developed using Matlab toolboxes before it was implemented on the robot using the C++ language through the ROS packages structure and the open source libraries OpenNI and OpenCV.

In the following sections, the proposed model is going to be presented step by step, from static to dynamic modalities and from visual to audio cues.

B. *Step 1 – Data retrieval*

This step’s goal is to get the data from the sensors.

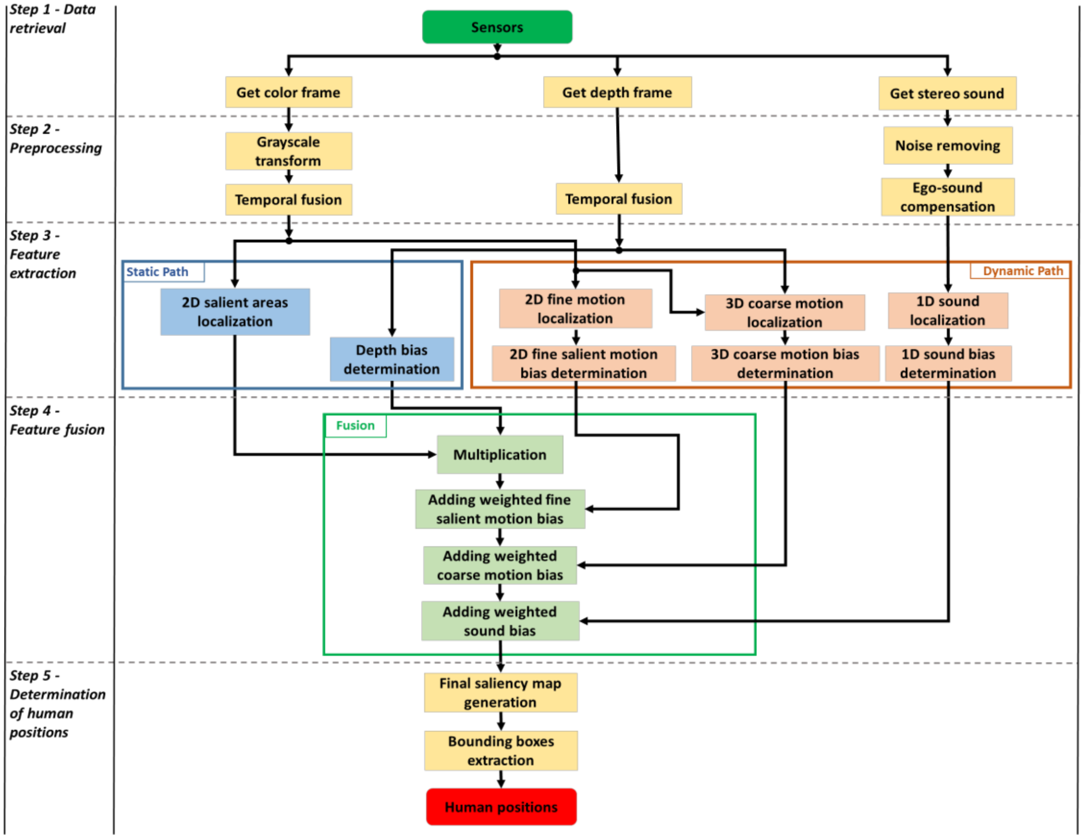
Step 1 - Data retrieval
Step 2 - Preprocessing
Step 3 - Feature extraction
Step 4 - Feature fusion
Step 5 - Determination of human positions
Sensors
Get color frame
Get depth frame
Get stereo sound
Grayscale transform
Temporal fusion
Temporal fusion
Noise removing
Ego-sound compensation
Static Path
2D salient areas localization
Depth bias determination
Dynamic Path
2D fine motion localization
3D coarse motion localization
1D sound localization
2D fine salient motion bias determination
3D coarse motion bias determination
1D sound bias determination
Fusion
Multiplication
Adding weighted fine salient motion bias
Adding weighted coarse motion bias
Adding weighted sound bias
Final saliency map generation
Bounding boxes extraction
Human positions

Figure 2. Flow chart of the proposed model.

As explained in Section I, color and depth images are both acquired at a speed of 30 FPS with a resolution of 640 by 480 pixels thanks to the RGB-D camera.

Stereo sound signals are discretized using a sample frequency of 44.1 kilohertz and bufferized in a one dimensional array.

C. Step 2 – Preprocessing

This step aims to reduce observation noises and computational time using simple but efficient operations. The following observations about the data have been made during this study:

- Color channels are not used in Step 3.
- Illumination variations generate noise on both color and depth images.
- Randomly located “holes” can be observed on depth images (i.e., not out of range areas that are considered as if they were out of range).
- Sound signals show an inconstant amplitude offset coming from the functionning robot’s system ego-sound.

The following operations have been realized. Their results are shown on Fig. 3.

First, the last retrieved color image is converted in grayscale, dividing by three times its computational cost.

Second, in order to improve the robustness of the images to noisy variations during time, a simple but efficient approach, driven by empirical considerations, has been chosen. It consists in blending the current image with the previous blended image. The resulting image is named reference image. Equation (1) describes this temporal fusion. This method has the advantage to take care of the hypothesis (2) made in section III because it does not consider explicit background information for illumination noise removing.

$$\boldsymbol{I_{ref}(t)} = \propto \times \boldsymbol{I(t)} + (\boldsymbol{1}-\propto) \times \boldsymbol{I(t-1)} \qquad (1)$$

In (1), α represents a parameter that can increase or decrease the importance of the previous images over time. The lower α is, the higher their importance is. Therefore, having a low α is important to consequently reduce noise variations over time, but it also gives less importance to the current image and tends to generate a less precisely localized motion. In the proposed model, α has been set to 0.8 for color (grayscaled) images and to 0.2 for depth images in order to smooth the depth holes while not having a high incident on the motion localization described in Sections III.D.3 and III.D.4.

Third, the noise from sound data is filtered with a low-pass 6th order Butterworth filter using a cut off frequency of

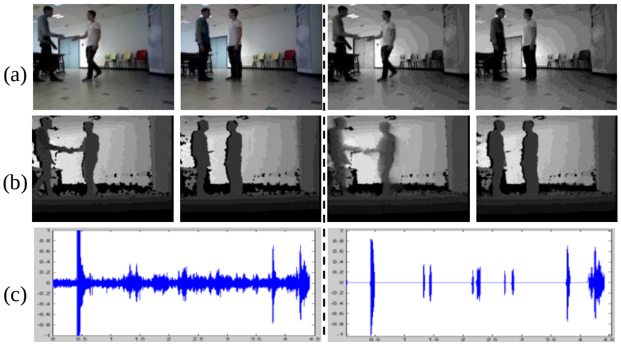

Figure 3. Example of data as obtained before (left) and after (right) the pre-process step. (a) color images; (b) depth images; (c) sound.

4 kilohertz designed on Matlab. It is then thresholded according to the intensity of its energy in order to avoid false detections coming from the ego-sound of the robot. The threshold used has been determined by analyzing energy values on 33 milliseconds records with and without external sound.

D. Step 3 – Feature extraction

This step represents the multiple processes that are applied to extract the interesting features used in the fusion step. It has been separated in two paths. The static path includes modalities that will always be detected. The dynamic path includes modalities that may be present at instant "t" but may be absent at instant "t+1". After each process of a same step, resulting images are normalized. The results are shown on Fig 4.

1) 2D salient areas localization: This process consists in applying a visual static saliency model to the previously grayscaled reference image. Like in [1], the model of [6] has been choosen for its efficency and its rapidity. Its principle is based on the spectral residual concept. The idea behind this is that salient areas on natural images (i.e., not artificial) may be considered as the less redundant ones. The method of [6] consists in applying a fast discrete two dimensional Fourier transform on a grayscale image of size 64 by 64 pixels. A logarithmic transform and a 3 by 3 mean filter are then applied to the amplitude spectrum of this image. The spectral residual spectrum is obtained via the substraction of the mean logarithmic representation with the original amplitude spectrum. The inverse fast discrete 2D Fourier transform is then used with the spectral residual spectrum instead of the amplitude spectrum. The resulting image is filtered with a 7 by 7 gaussian filter in order to obtain a 2D saliency map representing gaussian salient areas.

2) Depth bias determination: The goal of this process is to make the depth reference image represent closest values with the highest intensities in order to use the depth bias concept of [5] in the fusion step. It provides the robot with a more human-like perception model. It also helps to consider the hypothesis (4) described in Section III. First, the image is subsampled to a resolution of 64 by 64 pixels in order to be spatially equivalent to the saliency map. Its intensity values are then inversed. Values that are out of the sensor's range are set to zero in order to represent the absence of information. Finally, a closing operator with 3 by 3 rectangular structuring element is applied in order to remove the small detected holes.

3) 2D fine motion localization and fine salient motion bias determination: These processes aim to detect motion between consecutive reference images and to represent the saliency levels of the moving areas. This motion is considered as fine because it is detected on full resolution images. It is thus considered as able to capture motion having both small and high amplitudes. Since this process needs to be fast, a well known mean of the absolute difference operation is used via (2). The mean filter helps to remove false and small detected moving areas. Its size is of N by N pixels. N is equal to 7 in this model.

$$I_{diff}(x,y,t) = \frac{1}{N^2} \sum_{-\frac{N-1}{2}}^{\frac{N-1}{2}} |\, I(x+i, y+j, t) - I(x+i, y+j, t-1) \,| \qquad (2)$$

The process described in Section III.D.1. is then applied on the resulting difference image. The obtained result represents the detected motion through a salient gaussian areas representation: the 2D fine salient motion bias.

4) 3D coarse motion localization and coarse motion bias determination: These processes aim to detect motion between consecutive reference images that have been subsampled to a resolution of 64 by 64 pixels. It has been supposed to be only able to detect motions having a high enough amplitude. This motion is thus considered only if a fine motion as been detected. Moreover, since we have access to both depth and color (grayscaled) images, the coarse motion is detected on both in order to provide a more robust motion localization with the idea of a three dimensionnal motion. First, (2) is applied on the subsampled reference images. Second, an additive mean of the two difference images is realized in order to combine both motion representations. The result is binarized keeping only pixels with an intensity higher than 60% of the maximal possible value. The binarized areas are named blobs. Only blobs containing at least 40 pixels are considered as true moving areas. Their centroids are found and convoluted with a vertical gaussian whose size depends of the mean depth value obtained on a 3 by 3 area around the centroids. This representation is the 3D coarse motion bias.

5) 1D sound localization and sound bias determination: The sound is localized on the horizontal dimension using the cross correlation and the Interaural Time Difference (ITD) of [8]. First, a sound buffer of 33 milliseconds is retrieved from the two microphones. The size of this buffer corresponds to the required time in order to get an image with the camera. A cross correlation between left and right sound components is applied using a moving window of 20 samples. The ITD is then determined. It gives the angle position of the detected sound in the robot coordinates. This angle is converted in the 64 by 64 image pixels coordinates and a vertical gaussian is set at the sound location according to [10]. This is the 1D sound bias.

E. Step 4 – Feature fusion

This step aims to obtain the final saliency map by successively combining the various detected features from step 3 according to the hypotheses developed in Section III.

First, saliency and depth bias are combined through an element by element multiplication. This operation has been chosen in order to only impact the already salient areas. Then, the non-zero dynamic biases are successively added to the result. These operations are weighted additions in order to modify the relative saliency levels between the areas of the previously obtained image. The weights for fine motion, coarse motion and sound biases are respectively 60%, 60% and 30% in order to give a lot of importance to motions. Since experiments have shown that sound cannot be localized as precisely as visual features and as it is added after the motions, its weight is lower than motions' ones.

F. Step 5 – Determination of human positions

Human positions are retrieved using bounding boxes generated from the final saliency map. Bounding boxes are localized over the areas with a final saliency intensity of at least 50% of the maximal possible intensity in order to eliminate outliers without advantaging precision nor recall.

In order to define whether a detected bounding box should be considered as a human position, the correlation between a human being presence and the detected dynamic modalities has been learned. These modalities are represented by the dynamic biases and are the only available information sources that can help to make a *decision* about the eventual detection of a human being. The dataset described in Section IV has been used. The results are shown on Tab. 1. True positives (TP) correspond to a modality detection while a human is present, and false positives (FP) correspond to a detection when no human is present. True negatives (TN) and false negatives (FN) are also represented. The total detected (TD) values indicate when a modality has been detected over all the images.

Since all the false positive rates are low, it has been decided to use a simple binary *decision* for this model: if at least a fine motion or a sound has been detected, then it means that a human has been detected (i.e., is present).

TABLE I. CORRELATIONS BETWEEN DETECTED MODALITIES AND A HUMAN BEING PRESENCE (34 VIDEOS, 4 SOUND RECORDS).

	Fine Motion (%)	Coarse Motion (%)	Sound (%)
TD	82.7	42.5	2.2
TP	80	42.4	2.2
FP	2.7	0.1	0
TN	11.5	14.0	24.0
FN	5.8	43.5	73.8

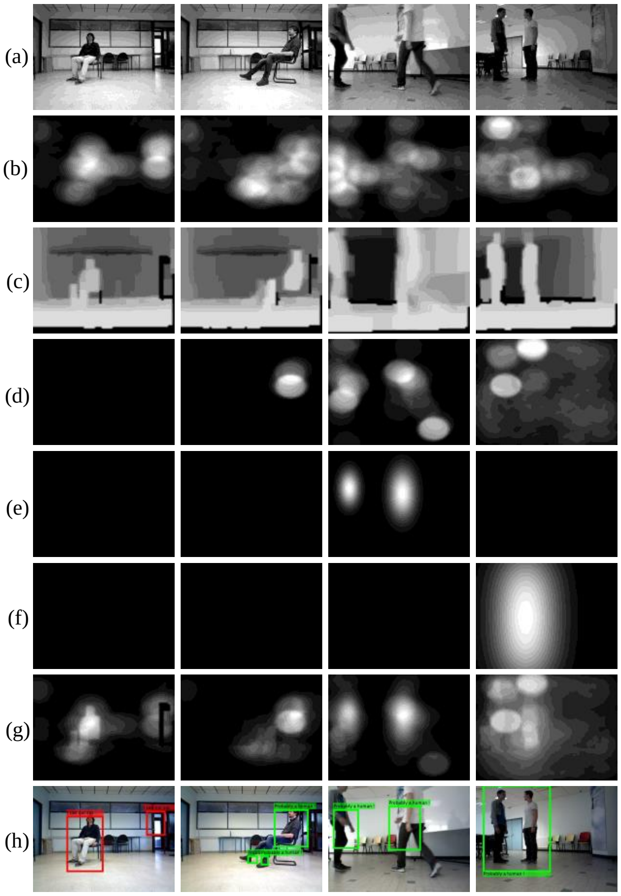

Figure 4. Example of results obtained by the model. Black images mean no information. (a) color (grayscaled) reference image; (b) salient areas; (c) depth bias; (d) fine motion bias; (e) coarse motion bias; (f) sound bias; (g) final saliency map; (h) bounding boxes. Red boxes indicate a lack of information for human detection, green boxes indicate probable human areas.

At this step, detected bounding boxes may be classified using their mean saliency values in order to determine which area is the most interesting. This operation may be useful in order to keep only one area to guide the robot's gaze towards the most interesting position.

IV. EVALUATION OF THE RESULTS

The results obtained by the proposed model have been evaluated on a dataset that has been acquired with the robot in two different rooms. It is made of 34 videos with durations between 7 and 20 seconds for a total of 10312 images. Only 4 videos also include sound data. At least one human is present on a part of each video. This human may be moving, sitting, standing or talking at any distance from the robot, but he is not necessarily always in the field of vision of the camera. The videos have been manually annotated with the frame ranges on which humans are present. For 10 of these videos (2 with sound), annotations also include the bounding box locations (ground truth) corresponding to the human 2D positions.

TABLE II. MEAN PRECISION, RECALL AND F-MEASURE WITH BOUNDING BOXES AFTER EACH FUSION (+). DETAILS FOR 2 VIDEOS WITH SOUND (W/+-) ARE PROVIDED. HIGHEST VALUES ARE IN GREEN, SECOND HIGHEST VALUES ARE IN BLUE. *DECISION* KEEPS ONLY BONDING BOXES WHEN HUMANS SHOULD HAVE BEEN DETECTED.

	Precision (%)		**Recall (%)**		**F-measure (%)**	
***Decision*?**	**yes**	**no**	**yes**	**no**	**yes**	**no**
Saliency	28	22	35	29	32	26
+ Depth	44	35	42	35	43	35
+ Fine Motion	**72**	56	**60**	50	**66**	53
+ Coarse Motion	**77**	60	**55**	46	**66**	53
w/- Sound	**79**	73	**44**	39	**57**	51
w/+ Sound	**76**	70	**49**	44	**59**	54

First, the resulting bounding box locations from step 5 have been compared with the ground truth using Matlab. The comparison method was to determine the precision, the recall and the F-measure between the bounding box areas obtained by the model and the ground truth. It has been determined that the model is able to generate a mean F-measure of 66% with a precision of 77% for human localization. The detailed results are shown on Tab. 2.

The following are a detailed explanation of these results. First, depth helps to increase both the recall and the precision of the human localization generated by the saliency. This corresponds to the fact that when humans are close to the robot, it is difficult to define very salient areas because humans are recovering a large amount of the image, it gives them an important spatial redundancy and induces difficulties for the method of [6]. Second, the hypotheses that have been made about the dynamic data are confirmed: they greatly improve the human localization. It is interesting to observe that the coarse motion bias does not improve the F-measure but the precision, and that the sound bias improves both the recall and the F-measure. Moreover, these results do not support the fact that humans may be detected even in the absence of dynamic data thanks to saliency. This means that one should use a specific detector on the detected areas in order to characterize them. In that case, the specific detector would not be used as an input of the model like in [12], but like a final recognition step.

Third, since a fast model was desired through the hypothesis (5) detailed in Section III, the computational time of the proposed model has been evaluated at different instants in time after the algorithm has been adapted in C++. The robot is able to process incoming flow at a mean speed of 70 FPS, which is twice more than required to process every frame using the Asus Xtion Pro Live RGB-D camera.

V. CONCLUSION AND FUTURE WORK

In this paper, an original approach to detect and localize human beings using audiovisual attention's concepts on an indoor companion robot has been presented. It is able to detect and localize humans at 70 FPS with a mean F-measure of 66% and a precision of 77% using bounding boxes on a stationary robot.

Since the proposed model uses motion and sound localization, future work will focus on studying the effect of ego motion compensation on this model using visual and non-visual odometry state-of-the-art methods. Adding a supplementary step in order to characterize detected areas while no dynamic data is available will also be studied. The adaptation of the proposed model to other depth sensors will be considered in order to make it suitable for an outdoor use.

AKNOWLEDGEMENT

This work has been partially supported by the LabEx PERSYVAL-lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir.

REFERENCES

[1] S. Anwar, Q. Zhao, S. I. Khan, F. Manzoor, and N. Qadeer, "Spectral saliency model for an appearance only SLAM in an indoor environment," 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, Islamabad, pp. 118-125, Jan. 2014.

[2] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," IEEE Trans. Pattern Analysis and Machine Intelligence vol. 35, no. 1, pp. 185-206, Jan. 2013.

[3] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient Object Detection: A Benchmark," in IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5706-5722, Dec. 2015.

[4] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," Journal of Vision, vol. 14, no. 8, pp. 1–17, 2014.

[5] J. Gautier and O. Le Meur, "A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions," Cognitive Computation, Springer, 4 (2), pp.141-156, 2012.

[6] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[7] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[8] Q. Labourey, O. Aycard, D. Pellerin, and M. Rombaut, "Audiovisual data fusion for successive speakers tracking," Computer Vision Theory and Applications (VISAPP), International Conference on, Lisbon, Portugal, pp. 696-701, 2014.

[9] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modeling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos," Int. J. Computer Vision, vol. 82, pp. 231-243, 2009.

[10] S. Ramenahalli et al., "Audio-visual saliency map: Overview, basic models and hardware implementation," Information Sciences and Systems (CISS), 47th Annual Conference on, Baltimore, MD, pp. 1-6, 2013.

[11] J. Ruesch et al., "Multimodal Saliency-Based Bottom-Up Attention, A Framework for the Humanoid Robot iCub," IEEE International Conference on Robotics and Automation, Pasadena, pp. 962-967, 2008.

[12] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot," IEEE Transactions on Human-Machine Systems, vol. 44, no. 2, pp. 157–168, 2014.