On Field Gesture-based Robot-to-robot Communication with NAO Soccer Players

Valerio Di Giambattista[1], Mulham Fawakherji[1], Vincenzo Suriani[1], Domenico D. Bloisi[2], and Daniele Nardi[1]

[1] Dept. of Computer, Control and Management Engineering, Sapienza University of Rome, via Ariosto 25 - 00184 Rome, Italy `nardi@diag.uniroma1.it`

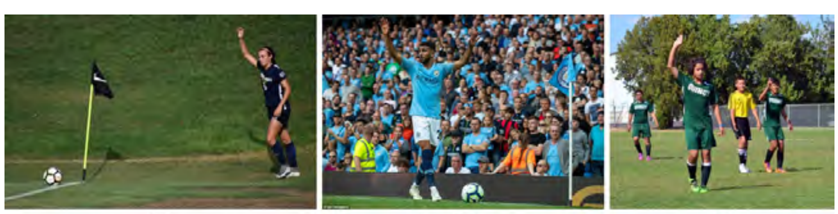
[2] Dept. of Mathematics, Computer Science, and Economics, University of Basilicata, viale dell'Ateneo Lucano, 10 - 85100 Potenza, Italy `domenico.bloisi@unibas.it`

Abstract. Gesture-based communication is commonly used by soccer players during matches to exchange information with teammates. Among the possible forms of gesture-based interaction, hand signals are the most used. In this paper, we present a deep learning method for recognizing robot-to-robot hand signals exchanged during a soccer game. A neural network for estimating human body, face, hands, and foot position has been adapted for the application in the robot soccer scenario. Quantitative experiments carried out on NAO V6 robots demonstrate the effectiveness of the proposed approach. Source code and data used in this work are made publicly available for the community.

Keywords: Communication protocols · Team coordination methods · Neural systems and deep learning.

1 Introduction

Professional and amateur soccer players commonly use gesture-based communications during games. For example, players use arms to call plays on corner kicks signalling where the ball is heading to (one hand normally means the front post, two hands means the far post — see Fig. 1a and Fig. 1b). Also holding up an arm while shouting for the ball is a common way to get a teammate's attention (Fig. 1c).



(a) (b) (c)

Fig. 1. Hand gesture during soccer games. Images from a) http://thenews.org b) https://www.dailymail.co.uk c) https://www.gftskills.com/soccer-sign-language

In this paper, we describe a deep learning approach for robot-to-robot gesture-based communication to be used during RoboCup soccer matches as in human's soccer. Our goal is to meet the intentions of the RoboCup Standard Platform League (SPL) committee to achieve more and more realistic games. Using gesture recognition is a possible way to deal with *i*) the recent introduction of free and corner kicks and *ii*) the limitation of wireless communications (no more than 1 message per second allowed). SPL at RoboCup 2019 permits the use of the V6 version of the NAO humanoid robot manufactured by SoftBank Robotics that is equipped with a more powerful motherboard than previous versions. This increase in available computational power allows to use deep learning frameworks, such as TensorFlow, Keras or PyTorch, that were unavailable in the past years due to hardware limitations.

The main contributions of this work are:

- The description of a pipeline for estimating the position of body keypoints on the NAO robot.
- The definition of a gesture-based message protocol for robot-to-robot communication inspired to the one used in human's soccer.
- The release of the source code of the proposed approach together with the data used for evaluating it.

The remainder of the paper is structured as follows. Related work is discussed in Section 2. The proposed method is presented in Section 3. Experimental evaluation is shown in Section 4. Finally, conclusions are drawn in Section 5.

2 Related Work

Developing human-like robot behaviours is a key aspect for dealing with the RoboCup 2050's challenge, consisting in creating a team of fully autonomous humanoid robot soccer players able to win a soccer game complying with the rules of FIFA against the winner of the World Cup. The actual trend in the RoboCup competitions is to rely less on WiFi communication. In particular, SPL decided to test audio communications by promoting specific technical challenges. However, audio communications suffer a lot from interference and are not robust to noise. Thus, a pure audio approach seems not sufficient to cope with the 2050's challenge.

In 2014, SPL added a technical challenge, called Drop-in Player Competition, where robot originating from different teams and with different software had to play together. This serves as a testbed for cooperation without pre-coordination (see [6]). Pennisi et al. [12] presented an open source framework to extract the orientation of the NAO robots on the field in the same year. The extraction were carried out off the NAO board using external RGB-D sensors. More recently, extraction of NAO orientations in SPL field over short and medium distances has been carried out on the robot hardware both using a domain-specific approach [11] and a Convolutional Neural Network (CNN) [10].

Human-robot interaction using gesture recognition is an active research field. For example, [5], [13]. In assistive scenarios, human-robot interaction can be performed with NAO robot platform using gesture communication [2]. In [8], a framework to track hand gestures has been proposed to interact with the same robot platform.

As a difference with existing work on human-robot interaction, we apply gesture based communication for exchanging messages between robots. Among the CNNs based approaches for human posture recognition, OpenPose [4] demonstrated to work well in a variety of scenarios. However, before the release of V6, NAO robots were too limited in computation power to perform CNN computation using general deep learning frameworks [1]. Inspired by human soccer, we propose in this paper a visual communication protocols based on robot posture. The use of postures to the robot-to-robot signal exchange problem is suitable for intention communication in mixed team competitions and in upcoming RoboCup scenarios such as corner kicks.

3 Methods

Fig. 2 shows the three main steps of our pipeline. The raw RGB image coming from the NAO top camera is transformed into the HSV color space and it is processed to find the 2D locations of anatomical keypoints for each robot in the image. The pose of the robot is inferred from the list of detected keypoints.

To extract the keypoints we use the open-source library OpenPose[3] proposed by Cao et al. [3]. OpenPose exploits Part Affinity Fields (PAFs) for multi-robot pose estimation. PAFs are sets of 2D vector fields that encode the location and orientation of limbs over the image domain. Even if this approach is particularly cost effective, it can achieve high-quality results. First, a feedforward network predicts a set of 2D confidence maps S of body part locations and a set of 2D vector fields L of part affinities, which encode the degree of association between

[3] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

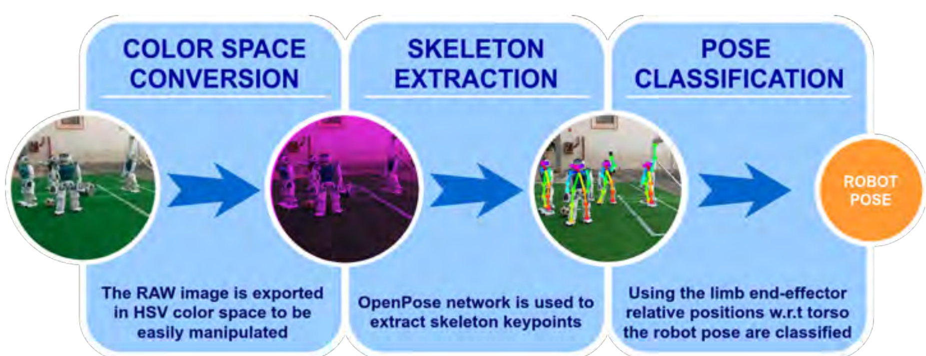


Fig. 2. Our pipeline for robot pose classification.

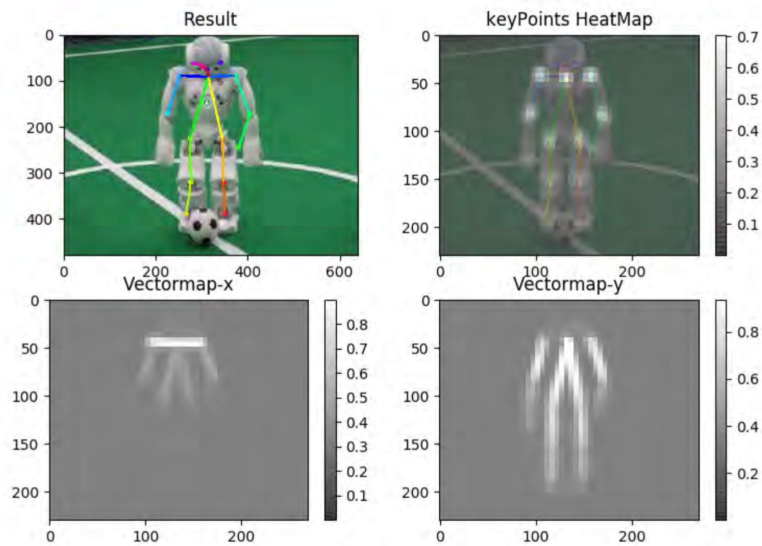


Fig. 3. Robot model.

parts. In our pipeline, when the skeleton is extracted from the image, it is classified in one of the four possible poses of interest, namely one arm raised, two arms raised, one arm raised to the side, and no arms raised. Once the current pose has been classified, it is possible to associate a message to the pose for non-verbal coordination.

Robot Model. The skeleton structure used in OpenPose for multi-person 2D pose detection can be used also in the case of NAOs, being them humanoid robots. Fig. 3 shows an example of the application of the OpenPose functions on an image containing a single NAO. The following keypoints are defined: $Shoulder$, $Elbow$, Hip, $Knee$, Eye, $Wrist$ (left and right for all this parts) and the $Neck$. A dataset of images was created by collecting images acquired with the NAO top camera during outdoor tests simulating corner kick game situations and the position of the above listed keypoints were annotated in accordance of the standard COCO-dataset notation used by OpenPose.

Network Architecture. The original network was modified to increase the confidence range of robot recognition. In this way, the individual mechanical parts of the robot are recognized through the keypoints and then correctly labeled. The network iteratively predicts affinity fields that encode part-to-part association and detection confidence maps. The iterative prediction architecture treats the predictions over successive stages, t ϵ $(1, ..., T)$, with an intermediate supervision at each stage. In our model the receptive field is preserved, while the computation is composed by three consecutive kernels and the output of each one of them is concatenated. This particular scheme allows to obtain accurate results and to deal with simultaneous detection and association.

We have used an auxiliary CNN model, build from the first 10 layers of VGG-19, to analyze the images taken by NAO robot's camera and extract a features map denoted as F. This features map F is processed by the first stage t_1, to generate a set of part affinity fields (PAFs) following $L_1 = \psi_1(\mathrm{F})$, where ψ_1 refers to the CNNs for inference at network's stage one. The output from this stage and and their predictions concatenated with the initial F are used as input for the next stage (as features), to produce refined predictions. This process is repeated on every stage, with T_P iterations where T_P represent the total number of PAF stages. The process is repeated to detect the confidence maps, starting from the updated PAF prediction. In this way, the computation time is reduced and the prediction of confidence map, are done on top of the latest and most refined PAF predictions. At the end of each stage we applied a loss function, to iteratively guide the network's PAFs predictions of body-parts, in the first branch and confidence maps in the second branch.

Fine tuning. To avoid a full training stage, we decided to carry out a fine-tuning on the models already available for people detection to obtain recognition even for NAO robots.

Implementation details. OpenPose can run on different platforms and provides support for different hardware, such as OpenCL GPUs and CPU-only devices. The inference time of OpenPose outperforms all state-of-the-art methods with high-quality results, like Mask R-CNN [9] and Alpha-Pose [7] multi-person estimation libraries. To obtain better computational performance, instead of the original OpenPose implementation based on the Caffe framework, we have decided to use TF-Pose network[4], an implementation based on TensorFlow library.

The original network has been modified in order to obtain an optimal recognition, for this reason the parameters of threshold-part-confidence in the PAF module has been modified going to make a 10 % increase with regard to the upper-body detection and a 20% increase for the face detection. The network was trained with 640× 480 input images, this to ensure that the input images were consistent with the dataset of the models already processed, and once the new model was obtained, true inference tests were performed with several image-sizes in input for optimal performance, as shown in the next section.

4 Experimental Results

To test the performance of the pose recognition task on challenging data, we created a dataset from a game scenario containing images captured outdoor with natural light conditions. This means that different areas of the scene may be subject to high contrast and changes in the brightness values. Fig. 4 shows some image samples from the dataset, which is available from download at: `http://www.dis.uniroma1.it/~labrococo/?q=node/459` as part of the **SPQR NAO**

[4] `https://github.com/ildoonet/tf-pose-estimation`

Fig. 4. Examples of images from the dataset used for the experimental evaluation.

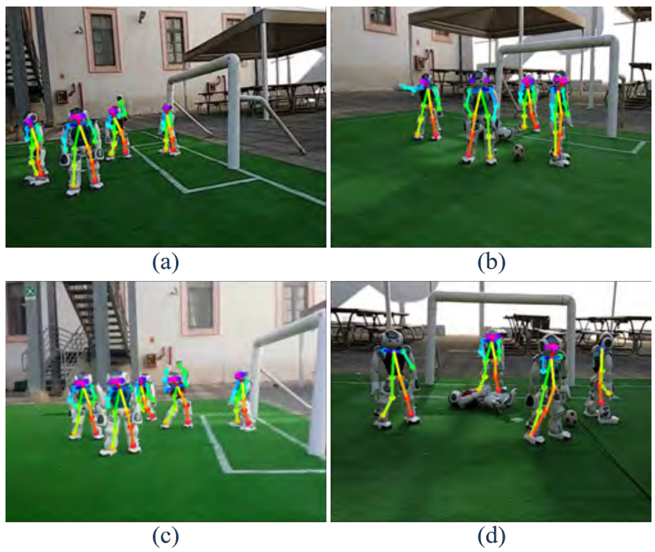


Fig. 5. Examples with successful classification. a) One arm raised. b) Arm to the side. c) Two arms raised. d) None – no message.

image dataset. The images were captured by using the top cameras of different players (i.e., NAO robots) in recreated game situations including corner and free kicks. Generated dataset contain 400 images, 100 images for each pose To the best of our knowledge, this is the first dataset for gesture recognition available in the RoboCup SPL specifically conceived for robot-to-robot gesture-based communication.

Quantitative Evaluation. Fig. 5 shows four examples of correctly classified robot poses. The quantitative evaluation of the gesture recognition accuracy is given in the form of the confusion matrix shown in Fig. 6. The gesture denoted as *two_arms_raised* is recognized with an accuracy of 89%, while the other gestures are recognized with lower accuracy. In particular, the lowest accuracy of 68% is obtained for predicting the gesture denoted as *arm_to_the_side*. This is mainly due to the overlap between body and arm of the NAO: 15% of the *one_arm_raised* pose is wrongly detected as *two_arms_raised* due to the similarity between the two poses.

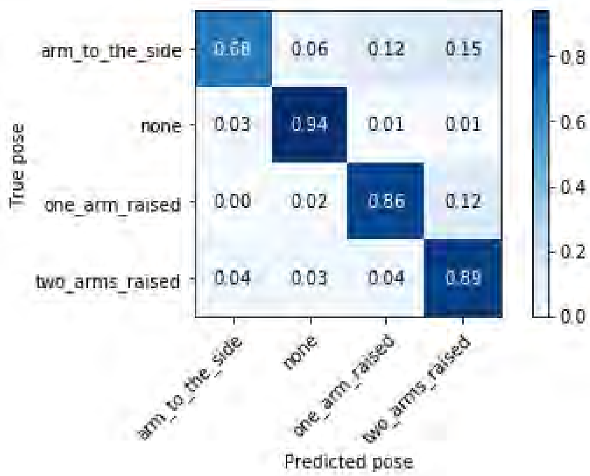


Fig. 6. Confusion matrix obtained using our robot gesture recognition approach

Table 1. Runtime performance of our gesture recognition approach on NAO V6 robot.

image size	CPU cores	CPU usage (avg.)	RAM usage	fps
232×217	4	67%	9%	2.1
320×240	4	85%	12%	1.2
432×368	4	96%	15%	0.88

Runtime Performance. NAO V6 robot is equipped with an Intel ATOM E3845 1.91 GHz quad core CPU 4GB RAM. To fully exploit the processing power provided by the CPU, the computational load has been spread over the four cores, obtaining the runtime performance shown in Table 1. Using an image size of 320×240 pixels it is possible to obtain a processing speed of about 1 frame per second (fps). It is worth noticing that, in our intention the gesture recognition module should be activated mainly in predefined game situations, such as corner kicks, thus a processing speed of 1 fps is feasible for possible use in RoboCup soccer matches.

5 Conclusion

We have presented a new approach for robot-to-robot gesture-based Communication based on a three-steps procedure to be used in the RoboCup SPL. Our aim is to propose a communication protocol similar to the one used in human's soccer games as a step towards the achievement of the Robocup 2050 challenge. In particular, the approach includes three main stages: 1) HSV color conversion to adjust illumination values; 2) Skeleton extraction to find robot's keypoints; 3) Pose classification to recognize predefined messages.

An important contribution of this work is the creation of a novel dataset, containing images captured from NAOs on a regular field taking three different

gestures. Quantitative experimental results demonstrate the effectiveness of the proposed approach. The source code and a tutorial to use it is available at the following link:
`https://github.com/SPQRTeam/Non-Verbal-Communication-With-NAO`.

As future work, we intend to implement the proposed approach on the GPU provided with the NAO V6 robot.

References

1. Albani, D., Youssef, A., Suriani, V., Nardi, D., Bloisi, D.D.: A deep learning approach for object recognition with nao soccer robots. In: RoboCup 2016: Robot World Cup XX. pp. 392–403 (2017)
2. Canal, G., Escalera, S., Angulo, C.: A real-time human-robot interaction system based on gestures for assistive scenarios. Computer Vision and Image Understanding **149**, 65–77 (2016)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 1302–1310 (2017)
5. Fujii, T., Lee, J.H., Okamoto, S.: Gesture recognition system for human-robot interaction and its application to robotic service task. In: Proc. of the International Multi-Conference of Engineers and Computer Scientists (IMECS). vol. 1 (2014)
6. Genter, K., Laue, T., Stone, P.: The robocup 2014 spl drop-in player competition: Encouraging teamwork without pre-coordination. In: AAMAS (2015)
7. H.-S.Fang, S. Xie, Y.W., C.Lu: Rmpe: Regional multi-person pose estimation. ICCV (2017)
8. Ju, Z., Ji, X., Li, J., Liu, H.: An integrative framework of human hand gesture segmentation for human–robot interaction. IEEE Systems Journal **11**(3), 1326–1336 (2017)
9. K.He, G.Gkioxari, P., R.Girshick: Mask r-cnn. ICCV (2017)
10. Leiva, F., Cruz, N., Bugueño, I., Ruiz-del-Solar, J.: Playing soccer without colors in the SPL: A convolutional neural network approach. CoRR **abs/1811.12493** (2018)
11. Mühlenbrock, A., Laue, T.: Vision-based orientation detection of humanoid soccer robots. In: Robot World Cup. pp. 204–215. Springer (2017)
12. Pennisi, A., Bloisi, D.D., Iocchi, L., Nardi, D.: Ground truth acquisition of humanoid soccer robot behaviour. In: RoboCup 2013: Robot World Cup XVII. pp. 560–567 (2014)
13. Sigalas, M., Baltzakis, H., Trahanias, P.: Gesture recognition based on arm tracking for human-robot interaction. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5424–5429. IEEE (2010)