Proxemics Toolkit For F-formation Patterns Detection

Mauricio Rivas, Paul Alvarez, Alfredo Barrientos, Miguel Cuadros
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
{u201511597, u201516795, pcsiabar, pcsimcua}@upc.edu.pe

***Abstract*—Interactions between people are of utmost magnitude for cross-device systems development. By using this kind of software, devices owned by those people end up interacting between themselves, and, therefore, making the system work. This work proposes to elaborate a toolkit that can detect and analyze those human interactions by using computer vision over videos showing them. All of these through the usage of 3D modeled test scenarios in addition to applying proxemics metrics and concepts of F-formations patterns so we can define them at various interaction types. To meet this goal, we used a previously trained human detection model in conjunction with two proposed concepts to estimate indispensable values: Distance between people, their body orientation, and relative position. To validate this tool, we tested it with a hundred test cases, each one having a set of different F-formation types so we could get the effectiveness of its detection functionality.**

I. Introduction

In recent years, interactions between people in collaborative environments have been more recurring. They usually have multiple devices interacting at the same time, for example, when sharing files, content, playing games, or edit collaborative content at a meeting [1]. These interactions could be improved by technology: people could actively use all available digital devices in a physical space, access to shared information could be improved, or even make presentation slides more dynamic as the expositor allow the audience to manipulate them in their own devices as wanted. All of this would be possible if there were innovative tools capable of performing more dynamic interactions between devices, specifically, through the use of cross-device interactions.

Projects, created to meet this need, face the issue of not having enough supportive resources since some of the major challenges in the making of cross-device systems is the lack of appropriate development tools, supportive infrastructures, and creation tools [1]. Considering this, we proposed a supportive tool for developers who are part of these projects.

Multiple people interactions can be computed by a video and there are, already proposed, various techniques to solve human detection; however, in this work, we focus on recognizing F-formation patterns, that can be interpreted as social or conversational groups, and the proxemics metrics, values related to the use of man's personal space, required for their detection, such as distance between people, body orientation of a person and the position of a person in a room, which will lead us to identify multiple human interactions that can take place on a closed space. Because of the different uses that can be given to this system, referring to the data that may be requested to it, it can be considered a toolkit.

Furthermore, usage of this tool can be of great value to secure social distancing while the global coronavirus pandemic is still a threaten. Because of this, one of the purposes of this work is to estimate the distance between people. Therefore, security camera videos could be analyzed in real-time to detect if two or more people have less than the recommended distance between them. This way, we could prevent the spread of the virus in a closed space.

As for the organization of sections for this paper, a review of the work related to the topic of the same will be made, as well as the description of the development of the toolkit, detailing its design, the creation of test scenarios to verify its operation, the realization of estimates for the distance between people and their body orientation. In addition, it will be informed about the process of validating the performance of the tool, and finally, the conclusions of the work and acknowledgments.

II. Related work

Research on academic works was done to demonstrate a theoretical and empirical basis. These papers, according to their traits, were divided into two groups. The first one deals with the human detection aspect by using computer vision. The other one focuses on F-formation patterns studies or projects created to recognize interactions between people of the same social or conversational group.

According to [2], Computer vision is the field resulting from the combination of image processing and pattern recognition. It is responsible for creating models and extract data and information from digital images. We elaborated the F-formation pattern detection by using video analysis, by using videos in which people can be seen interacting with each other. We present some works that make use of this area to fulfill their development:

In [3], the authors proposed an efficient human detection with the usage of deep learning and standardization of human-aware restrictions. They proposed to cascade the aggregate channel features (ACF) with a deep Convolutional Neural Network (CNN) to achieve fast and accurate Pedestrian Detection. They used a mixture of asymmetric Gaussian functions, to define the cost function associated with each constraint.

In [4], they proposed a social relation impression management scheme to protect relational privacy and to automatically recommend an appropriate photo-sharing policy to users that, additionally, measures a distance between user's faces within group photos by relying on photo metadata and face-detection results. These distances are then transformed

into relations using proxemics. Furthermore, they proposed a relation impression evaluation algorithm to evaluate and manage relational impressions.

The search for the other group of works, about f-formation detection, was carried out with the purpose of looking for the principles and techniques, applied in their detection models, that may serve as a guide to develop a practical toolkit for detecting these patterns. An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access [5], they also could be defined as a conversational group of people or some persons performing a group activity. They can be classified into different types according to some metrics, known as proxemics dimensions. Proxemics can be defined as the study of the nature, degree, and effect of the spatial separation individuals naturally maintain (as in various social and interpersonal situations) and of how this separation relates to environmental and cultural factors [6]. Works related to these patterns are described below.

In [7], they proposed a method to detect steading conversational groups to obtain an image that describes the displacement that these individuals would develop. They based their solution on the fuzzy relation theory and their clustering to acquire a better quantity of detected F-formations on a still image. Their proposal performance was evaluated and compared against other reported methods over two real-world databases. Their experimental results show the effectiveness of this work.

In [8], a novel framework for jointly estimating head, body orientations of targets, and conversational groups. Their algorithm employs body pose as the primary cue for F-formation estimation. To estimate body pose precisely, their learning framework limits the possible range of body orientations based on head posture and body orientation to jointly learn both. Additionally, it proposes an alternating optimization strategy to iteratively refine F-formations and pose estimates. Lastly, they proved the increased efficacy of joint interference over the state of the art via extensive experiments on three social datasets.

In [9], they explore mobile collocated proxemics interactions by observing F-formations to provide a better understanding of how people socially interact in non-traditional, non-structured, dynamic environments. This study was conducted as follows: All 12 members were required to observe and make notes of anything that seemed unusual or that was not previously stated in the aforementioned studies of F-formations in controlled settings. Such observations could include information about group size, movement patterns in an open space, the physical distance between people, or possible usage of devices. It was concluded that the most frequent F-formation was when people gathered in a semicircular layer, the size of those groups varied between 2 and 15 people, and that the sizes of those groups with their physical distribution were often altered as people moved from one place to another.

According to [10], detecting free-standing conversational groups or F-formations in surveillance videos introduces a new social representation for computing relations between individuals. Their proposed solution is based on fuzzy relations in which a membership function for computing the interception between person frustum distance relations. In surveillance video scenes, the presence of crowds grows over time. The analysis of these scenes is significant to prevent, predict and detect dangerous situations, such as riots, manifestations, terrorist acts, among others, through systematic observations. This analysis becomes a hard task for humans because psychological studies suggest that their perceptions become impacted when, in crowded settings, multiple scenes are analyzed simultaneously. The proposal uses aspects extracted from each video frame. These traits are used in the social representation step which generates a fuzzy matrix for each frame.

In [2], a method for the automatic detection of F-formation patterns in still images is proposed. Authors submit the following process: First, they furnish the rigorous definition of a group taking into consideration the background of the social sciences: this allows them to specify many kinds of groups. On top of this taxonomy, they present a detailed state of the art on the group detection algorithms. Then present a method for the automatic detection of groups in still images based on a graph-cut framework for clustering individuals; in particular, they can codify, in a computational sense, the sociological definition of an F-formation that is very useful to encode a group having only proxemics information: position and body orientation of people. Better results were obtained in the recognition of different types of F-formation patterns compared to other algorithms that achieve the same purpose.

In [11], the authors introduced two novel algorithms for detecting groups of people who are standing or freely moving in a crowded environment. The first algorithm, called the “Link method”, uses a learning and forgetting strategy for modeling dynamics of proxemics between individuals. The second algorithm, called the “Interpersonal Synchrony Method”, explicitly adopts interpersonal synchrony to refine clusters of persons detected by combining proxemics and 2D field of view individuals. The proposed algorithms were evaluated on both simulated and real-world video sequences from state-of-the-art databases.

From the works described above, we can affirm that our project simplifies the necessary development for achieving detection of F-formation patterns, through mathematical models and by using a pre-trained human detection model (OpenPose). The main disadvantage of this project, compared to the previous ones, is the quality of the tests carried out on the toolkit because of being unrealizable to carry out face-to-face meetings with test subjects due to the development of this work during one of the highest points of the current pandemic. However, the quality of the detection model used in the validation of the toolkit phase, makes it possible to obtain better results when applied in real-life scenarios.

III. DEVELOPMENT

For the development of this project, we designed a toolkit based on a referential architecture, described later in the paper. By using 3D models, we developed an estimation of the distance between people and their body orientation. Finally, for detecting F-formation patterns, we set special conditions, on the physical position and body orientation, to measure the detection precision of the tool.

A. Toolkit design

It was planned to develop a toolkit of video analysis within a local deployment to avoid an overload on an external server when it receives multiple requests, this decision was made because of human detection in addition to the estimation of some metrics, necessary for F-formation detection, will consume large amounts of computational resources. It would be used in a 4 by 4 room, using a webcam, located in the upper center of one of the walls, with a focal length that allows covering the entire image of the room.

To start executing the software, the user must enter the values that he wants to obtain (distance, body orientation, and position) as the input data and set the connection with the camera. Using real-time video streaming, the toolkit begins to analyze the live images of what happens in the room, and, at the moment that an F-formation is detected, the user would be notified, and the data that he requested, in the beginning, would be provided.

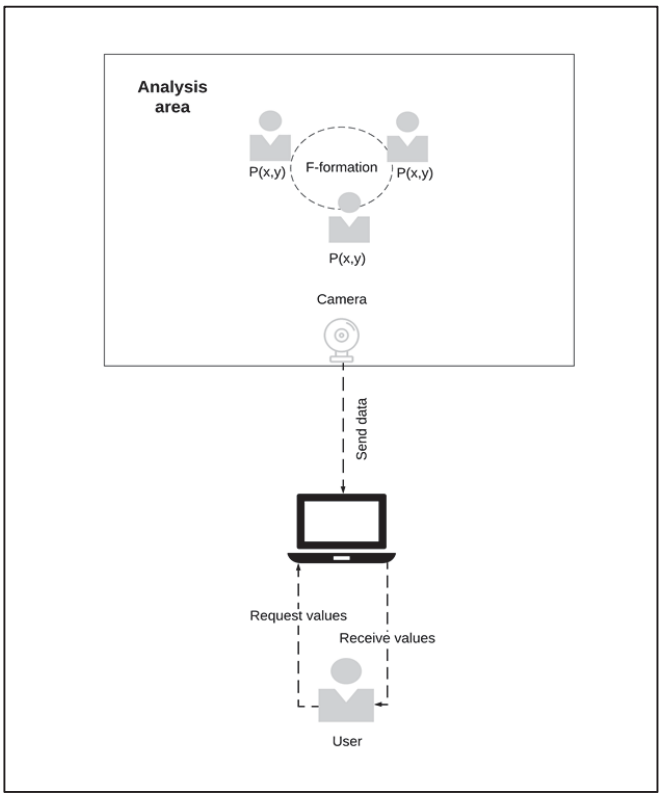

Fig. 1. System referential architecture

Since this work was done during the coronavirus pandemic, we couldn't test the toolkit with real people, instead, we created test scenarios to simulate the situations we needed to validate its performance, and we focused on the detection model development and its functionalities. However, toolkit operation in a real environment remains the same as described above.

B. Creation of test scenarios

Because of the needing for the toolkit to be used on videos in which people are moving inside a predefined closed space, 3D animations made in Blender, a software for tridimensional graphics creation, were used for testing the toolkit. This way, we could validate the results of the project by simulating multiple scenarios where simulated people are moving around inside a room and rotating their bodies, in order to create F-formation patterns.

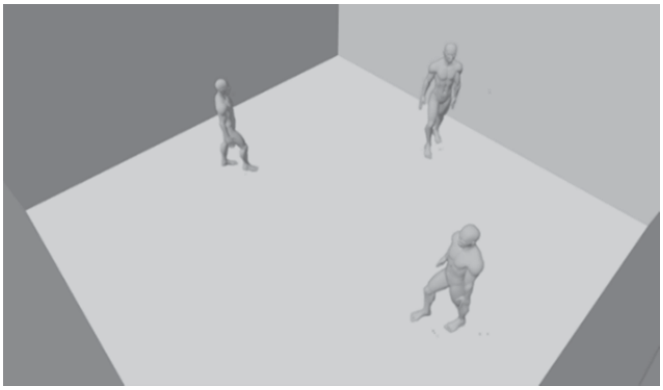

Fig. 2. Test scenario used to validate the toolkit

C. Estimation of the distance between people

To detect f-formation patterns it is imperative to know if some of the detected people are at a short distance from another one. This is in order to consider them as one of the participants of an F-formation pattern. Therefore, we decided to use a mathematician formula known as Euclidean distance to compute this metric applied to every pair of persons detected by the toolkit.

According to [12], Euclidean distance is defined as an application of the Pythagorean theorem on right triangles where the Euclidean distance is the length of the hypotenuse of the right triangle formed by each point and the projected vectors over the directed axis at the hypotenuse level.

In a coordinate plane drawn from the dimensions of the input video. The points A and B in which A = (x_A; y_A) and B = (x_B; y_B) represent the position of two people in the room, we define Euclidean distance between those two points as:

$$distance(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

where, x_A and x_B are the values of the "x" axis, in pixels, of the points A and B, respectively, as y_A and y_B for the "y" axis.

A distance extrapolation was done to convert the number of pixels, between each detected person, in meters. This was possible with an image processing of the dimensions of the 3D modeled room used for the scenarios. Hence, by using OpenCV for detecting the contrast in colors between the floor and walls of the room, we were able to recognize each corner of the floor as a coordinate pair (x, y). The floor image was segmented into ten sections in which their height is equal, but their width value may increase if the section is closer to the camera. Therefore, we could get different pixel-to-meter conversion ratios, by dividing the image width of each sector into the real width of the floor, to obtain a more realistic distance reference of the horizontal space in the room, respecting how far is the camera from a certain point in the floor, with an error margin of 5%.

Fig. 3. The computing of a room image to extrapolate distances

The person's feet were taken as a reference point to obtain their physical position in the room and, by applying the Euclidean distance formula, the estimation of the distance between each person was achieved.

D. Estimation of people body orientation

To estimate the body orientation angle for each detected person we applied the following formula:

$$angle(A,B) = arctg\left(\frac{y_B - y_A}{x_B - x_A}\right) \times \frac{180}{\pi}$$

where A and B are the points belonging to the right and left shoulders of a detected person. As a consequence, x_A and x_B represent the values, in pixels, of the "x" axis of both points respectively, as y_A and y_B in the "y" axis. By calculating the arctangent of the slope of those points, we could find the inclination angle, on radians, in respect of a horizontal axis. Thus, by multiplying this value by 180 then dividing it by pi, we can get its converted value to sexagesimal degrees.

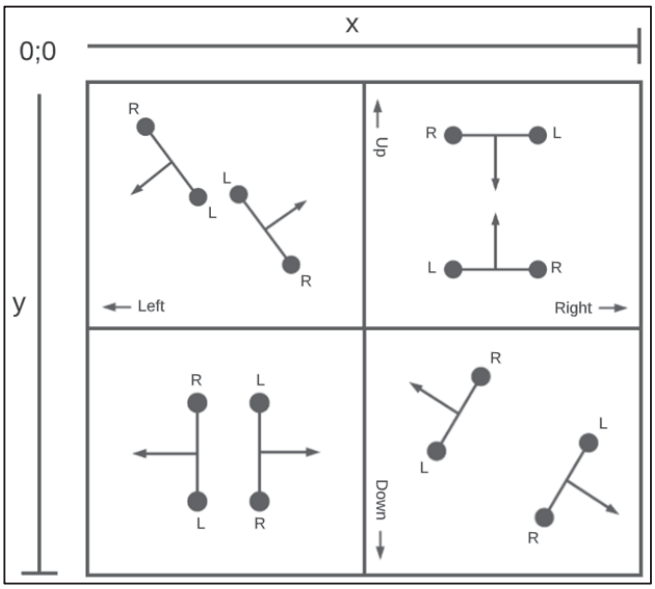

Fig. 4. The reference model for the estimation of body orientation angles

E. F-formation pattern detection

From the fact that two of the detected people are at up to one meter and the difference between their body orientation angle regarding the camera, we defined angle thresholds, for the body orientation of a person, that establishes the type of F-formation detected. We considered three formation types: Face-to-face, side-to-side, and "L" type.

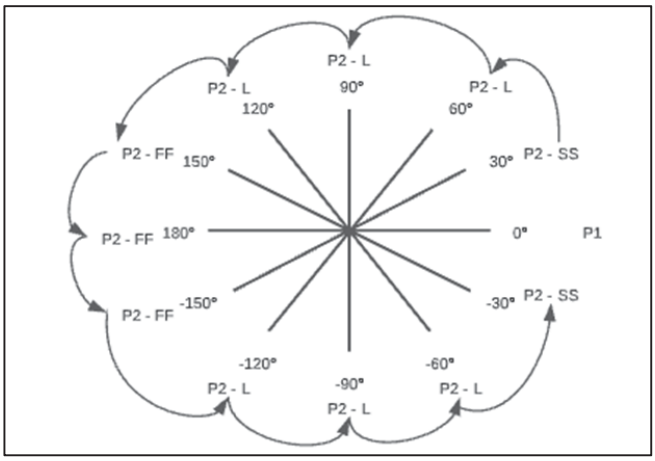

Fig. 5. Definition of F-formation types regarding the body orientation angle variance between two people

In Fig. 5 we could perceive the definition of those thresholds to each formation type. "P1" and "P2" refer to two people, "SS" to side-to-side type, "L" to "L" type, and "FF" to face-to-face type. It also shows how the type of formation varies according to the angle formed by the body orientations in reference to a horizontal axis.

Each pair of people detected by the toolkit were analyzed to define if they are part of an F-formation pattern. If it detects that a person is part of more than one formation at the same time, we unify those into one bigger formation that integrates all participants. Considering this, we also add two new F-formation types, “three members” and “four members”, regarding the number of persons in the group.

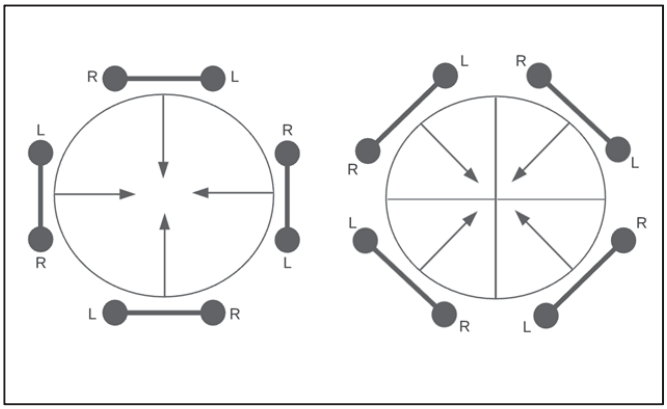

Fig. 6. The reference model for F-formation patterns detection

F. Toolkit elaboration

Regarding people detection through cameras, we agreed upon using OpenPose, a human detection model, as the base project because of its multiple benefits for the project:

- Allowing us to detect a group of people on a real-time video or a previously recorded one.
- The capability to recognize specific body key points such as the nose, neck, shoulders, wrists, elbows, hip, knees, ankle, eyes, ears, and thumbs.
- Ability to retrieve a coordinate (x,y) using the top left corner of the video/image, inserted as the input data, as the starting point (0,0).
- Drawing skeletons, built by these key points, on each detected person.

Considering the actual use of the toolkit, it is important to emphasize that the base detection model for this project is designed to detect human bodies, so the possible operation of this toolkit with real people should have a performance equal to or better than what is shown in this work. Thus, we can compensate for the results, in case we expect better figures than those obtained in the validation of the toolkit.

The OpenPose project needs certain previous steps to be able to start executing and to be modifiable. The first steps consisted in the installation of the libraries necessary for its first execution, among them are "argparse", "dill", "fire", "matplotlib", "numba", "psutil", "pycocotools", "requests", "scikit-image", "scipy", "slidingwindow" and " tqdm ", but mainly Opencv and TensorFlow were used, in these you have to take into account the installed versions, you can choose to use a more updated version, but the base code does not support this, so you need to make various changes through the files that compose it. Another alternative can be to use the recommended, by the project creators, outdated versions, so that there are no major complications on the base code.

Fig. 7. Shoulder position estimation by OpenPose from a detected person

Of the OpenPose project, the file that is responsible for detecting human bodies, called estimator.py, was modified. In this file, we store the values for the positions of the shoulders of each detected person, and here we also calculate the center point of the segment formed by those coordinates, as a reference point for estimating the body orientation. We also perform detection iterations between the people detected, throughout the input video playback, to check how their positions and orientations change while they are moving across the room.

By the aforementioned characteristics, the positions of the necessary body key points (shoulders and feet), of each person in the scenarios, were obtained. Making it possible to calculate the distance between people and the body orientation of a person. Therefore, by applying the earlier explained concepts, we did get the required metrics to complete our main goal: recognize an F-formation pattern.

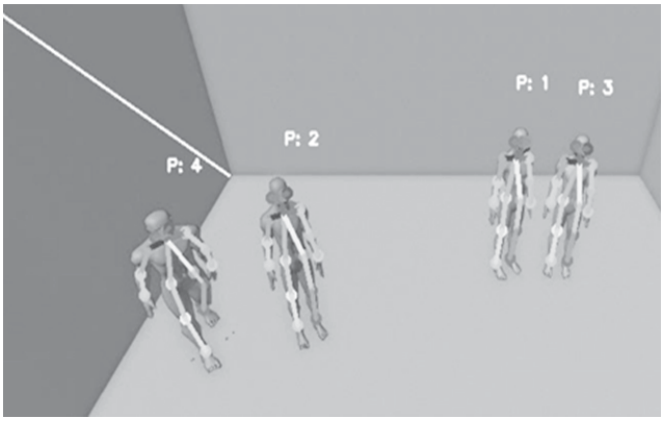

Fig. 8. Toolkit analysis sample on a test scenario

As a result of the toolkit analysis, the console output notifies us when an F-formation pattern arises in the input video and what type of formation it is. The output also includes the position, in "x" and "y" coordinates, and the body orientation angle of each member of the F-formation detected.

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Person 1  and Person 3 are doing an f-formation
- Person  1 : angle =  0
- Person  3 : angle =  8
- Person  1 : coordinate center =  [656 218]
- Person  3 : coordinate center =  [717 222]
f-formation de tipo SIDE TO SIDE
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Person 4  and Person 2 are doing an f-formation
- Person  4 : angle =  -21
- Person  2 : angle =  -14
- Person  4 : coordinate center =  [308 283]
- Person  2 : coordinate center =  [429 266]
f-formation de tipo SIDE TO SIDE
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

Fig. 9. Outgoing output data as the toolkit is analyzing the input video

G. Validation

The validation goal of the current work consisted of the analysis of a hundred test cases. We defined a test case as the correct detection of an F-formation pattern, shown in a scenario. The 3D modeled people, used for the validation process, had different physical characteristics, like height or body build, to make it more similar to a real-life people. The movement of the persons was random, prioritizing the realization of the previously defined f-formation types.

TABLE I. RESULTS OF THE PROJECT VALIDATION ACCORDING TO EACH TYPE OF F-FORMATION PATTERN

F-formation type	Detected	Non detected
Four members	5	2
Three members	13	3
Face to face	20	5
"L" type	15	7
Side to side	27	3
Total	80	20

The toolkit managed to detect eighty formations out of a hundred. Through this process we could find some remarks about the detection effectiveness:

- This toolkit does not recognize people shorter than 1.45 m.
- The developed toolkit has better performance at detecting F-formation patterns of type Side to side.
- The "L" type F-formation pattern has the least cases of successful detections.
- When an F-formation pattern has more than two members, the toolkit tends to effectively detect it.

For a better understanding of the validation results, a stacked bar graph was drawn up, shown in Fig. 10, detailing the ups and downs of the toolkit performance in the test cases.

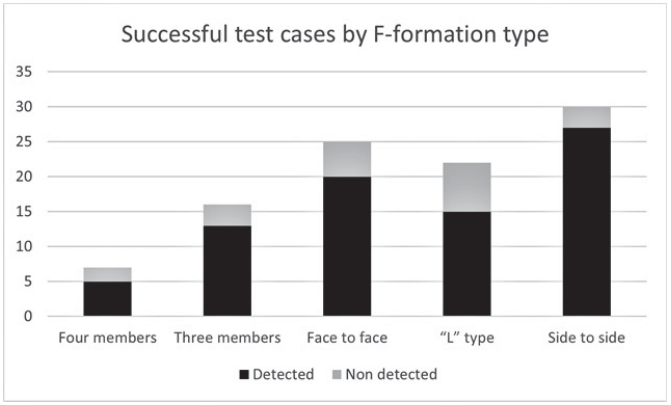

Fig. 10. Results stacked bar chart

IV. CONCLUSION

A toolkit for detecting F-formation patterns in closed spaces was developed, through the usage of computer vision. In addition to notifying the user when detection is done it specify into which category type the formation belongs. Also, it estimates proxemics metrics like the distance between people, body orientation, and the position of each detected person person inside the room. These functionalities expand the range of possibilities of use for this toolkit, being able to be useful for other projects besides cross-device systems.

Regarding the toolkit performance, positive results were obtained as 80% effectiveness was achieved by testing it on 3D modeled test scenarios. It should be emphasized that a detection model, designed for real people, was used, so its effectiveness, while being applied to real-life videos, would be higher.

Lastly, external projects in need of detecting physical interactions between people, as cross-device projects, can take advantage of this work and use it to facilitate and accelerate their development. This type of technology will be of great importance throughout the duration of the coronavirus pandemic, or in a future case that urgently requires social distancing.

V. ACKONWLEDGEMENT

We would like to especially thank the professors of the "Universidad peruana de ciencias aplicadas" who advised us from the conception phase of the project to obtaining its results, and the university itself for supporting us with the costs generated by the development process.

REFERENCES

[1] S. Houben, N. Marquardt, J. Vermeulen, C. Klokmose, J. Schoning, H. Reiterer and C. Holz, "Opportunities and challenges for cross-device interactions in the wild", *Interactions*, vol. 24, no. 5, 2017, pp. 58-63.

[2] F. Setti, C. Russell, C. Bassetti and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images", *PLoS ONE*, vol. 10, 2015.

[3] A. Mateus, D. Ribeiro, P. Miraldo and J. C. Nascimento, "Efficient and robust Pedestrian Detection using Deep Learning for Human-Aware Navigation", *Robotics and Autonomous Systems*, vol. 113, 2019, pp. 23-37.
[4] F. Li, Z. Sun, B. Niu, Y. Guo and Z. Liu, "SRIM Scheme: An Impression-Management Scheme for Privacy-Aware Photo-Sharing Users", *Engineering*, vol. 4, 2018, pp. 85-93.
[5] Profs.scienze.univr.it, F-formation discovery in static images, Web: http://profs.scienze.univr.it/cristanm/ssp/.
[6] Merriam-webster, Proxemics, Web: https://www.merriam-webster.com/dictionary/proxemics.
[7] E. Ferrera-Cedeño, N. Acosta-Mendoza and A. Gago-Alonso, "Detecting Steading Conversational Groups on an Still Image: A Single Relational Fuzzy Approach", 2019.
[8] J. Varadarajan, R. Subramanian, S. R. Bulò, N. Ahuja, O. Lanz and E. Ricci, "Joint Estimation of Human Pose and Conversational Groups from Social Scenes", *International Journal of Computer Vision*, vol. 126, 2018, pp. 410-429.
[9] A. Lucero and M. Serrano, "Towards proxemic mobile collocated interactions", *International Journal of Mobile Human Computer Interaction*, vol. 9, 2017, pp. 15-24.
[10] E. Ferrera-Cedenõ, N. Acosta-Mendoza, A. Gago-Alonso and E. Garciá-Reyes, "Detecting free standing conversational group in video using fuzzy relations", *Informatica (Netherlands)*, vol. 30, 2019, pp. 21-32.
[11] O. Islas, G. Varni, M. Andries, M. Chetouani, and R. Chatila, "Modeling the dynamics of individual behaviors for group detection in crowds using low-level features". *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 1104 -1111.
[12] Ecured, Distancia euclídea, Web: https://www.ecured.cu/Distancia_euclídea.