Multimedia Tools and Applications (2022) 81:73–94
https://doi.org/10.1007/s11042-021-11077-7

1191: SENTIENT MULTIMEDIA SYSTEMS

Enhancing spatial perception through sound: mapping human movements into MIDI

Bernardo Breve[1] · Stefano Cirillo[1] · Mariano Cuofano[2] · Domenico Desiato[1]

Received: 22 June 2020 / Revised: 3 February 2021 / Accepted: 11 May 2021 /
Published online: 18 June 2021

Abstract
Gestural expressiveness plays a fundamental role in the interaction with people, environments, animals, things, and so on. Thus, several emerging application domains would exploit the interpretation of movements to support their critical designing processes. To this end, new forms to express the people's perceptions could help their interpretation, like in the case of music. In this paper, we investigate the user's perception associated with the interpretation of sounds by highlighting how sounds can be exploited for helping users in adapting to a specific environment. We present a novel algorithm for mapping human movements into MIDI music. The algorithm has been implemented in a system that integrates a module for real-time tracking of movements through a sample based synthesizer using different types of filters to modulate frequencies. The system has been evaluated through a user study, in which several users have participated in a room experience, yielding significant results about their perceptions with respect to the environment they were immersed.

Keywords Movements tracking · MIDI sound · Synthesizer sounds

1 Introduction

Gestures, movements, and body languages represent a common way through which it is possible to mark verbal communication. Indeed, they emphasise the language by adding

✉ Stefano Cirillo
scirillo@unisa.it

Bernardo Breve
bbreve@unisa.it

Mariano Cuofano
mariano.cuofano@alumni.rca.ac.uk

Domenico Desiato
ddesiato@unisa.it

[1] Department of Computer Science, University of Salerno, Fisciano, SA, 84084, Italy

[2] MRes Architecture, Royal College of Art, South Kensington, London, UK

significant characteristics useful for communication purposes. More specifically, it is possible to capture details that allow to associate a particular gesture with an emotion. In fact, by analysing the interaction between two persons, it is possible to understand the feelings produced during their communication through their gestures.

In the state of the art, it is possible to find different tracking techniques for detecting human movements [4], and techniques to represent their semantics [1]. They have been mainly applied to scenarios in which it is necessary to support humans in the real-time interpretation of critical situations of daily life, such as in the context of video surveillance [7].

In the social context, an important role is played by the music. It is possible to define music as an art manifestation, since it consists of creating and producing sounds that are pleasant to the human ear. The music, in most cases, leads emotions in individuals who are providing and listening to it. Thus, also in this case, it is possible to associate a sentiment to a specific organisation of sounds. In fact, most of the time, people use music to regulate their emotions, this means that each individual through the listening of particular shades of sound modifies his/her emotional state or, more simply, a sound can create emotions such as happiness, sadness, gladness, and so on [10]. It represents a fascinating topic for researchers such as the ISMM Team at Ircam[1] that operates on interactive music systems, gesture and sound modelling, interactive music synthesis, gesture capture systems and motion interfaces.

In this paper, we exploit the combination of human movements and sound synthesis techniques to associate sounds to movements in the space. We aim to assess how sounds are beneficial for improving the user's perception when adapting to a specific environment and investigate how sounds affect this adaptation. More specifically, we propose a novel algorithm and a system for recognising human movements and translating them into MIDI music, with a sample-based synthesiser, which uses different types of filters to modulate frequencies. The resulting frequency after the modulation step is used to associate a specific set of sounds to each movement captured in real-time. In particular, to quickly map spatial coordinates of people in sound, we defined a layer-based module, which is able to uniquely identify MIDI notes and to manage their changes according to people's movements.

Moreover, we have also integrated a new module to map MIDI notes into Scientific pitch notation, which permits to represent the notes by using literal notation. This module allows the system to create playable music by exploiting the movements. To this end, it is possible to associate for each user a specific music sheet to provide a concrete representation of the movements into space.

The paper is organised as follows. In Section 2, we describe recent works concerning movements tracking techniques and their application into several fields. In Section 3, we introduce some preliminary concepts useful to understand our proposal. In Section 4, we explain our methodology, whereas its validation is discussed in Section 5. Finally, conclusions and future research directions are provided in Section 6.

2 Related work

In this section, we describe approaches and methodologies defined in the literature concerning the mapping of gesture recognition [3]. In particular, we focus on different application domains in which the recognition of gestures plays an important role [20].

[1] https://ismm.ircam.fr/

In the music field, interesting pioneering work is described in [19]. Authors present Conductor's Jacket, a wearable device that interprets physiological and gestural stimuli to apply them in a musical context. It uses sixteen sensors communicating with musical software by collecting data over different reliable channels and offering graphical feedback to control the mapped gestures. Instead, in [26] authors present MATRIX (Multipurpose Array of Tactile Rods for Interactive eXpression), a musical interface for music amateurs and professionals. It permits the usage of hands to control music by exploiting a 3-dimensional interface allowing the manipulation of traditional musical instruments in conjunction with it. The MATRIX interface manipulates the parameters of a synthesis engine or effect algorithm in real-time, responding to the performer's expressive gestures. In [9], authors illustrate a manifold interface that exploits three human performance abilities, to see, move, and listen for visualizing control parameter space the domain of which corresponds to the domain of input gestures. Their target is to extend gestures input to a 3D visual representation of the control space in which one can efficiently apply multiple parameter variation techniques by carrying out movements. In [8], the authors illustrate a real-time musical conducting gesture recognition system that supports music players in enhancing their performance. They used a single-depth camera to capture image inputs and to establish a real-time gesture recognition system. In the data mining field, one of the most relevant works is [25]. The authors propose an innovative framework for progressive mining and querying motion data exploiting information extracted from data relationships [6]. Another application field concerns the application of gesture recognition applied to video surveillance. In [24], authors propose an intelligent model that focuses their tracking on outdoor activities. In fact, the model allows them to detect and recognize activities in public environments by combining multiple video sources from several cameras. Although this model has proved to be efficient in the context of video surveillance, it uses a combination of video sources that could be not efficient for our goals. In [28], authors present an online approach to simultaneously detect 2D poses of multiple people in a video sequence. It exploits Part Affinity Field (PAF) representations designed for static images, and they propose an architecture that can encode Spatio Temporal Affinity Fields (STAF) across a video sequence. Although it has been proved to be efficient, especially in the context of sports, the main goal of the model is to track the poses of the people, to determine their behavior. In our model, it is not necessary to detect the poses of the people, but we focus on the position they occupy in space. This allows us to obtain an efficient model for our goals.

In [30], authors present a novel multi-person tracking system for crowd counting and normal/abnormal events detection in indoor/outdoor surveillance environments. They use two challenging video surveillance datasets, such as PETS2009 and UMN crowd analysis datasets, to demonstrate their proposed system's effectiveness, which achieved 88.7% and 95.5% of accuracy and detection rate, respectively. In [4] authors propose an algorithm for multi-person tracking in indoor surveillance systems based on a tracking-by-detection approach. They use Convolutional Neural Networks (CNNs) for detecting and tracking people. They also perform several experiments by tracking people in rapid-panic scenarios, achieving good performances in terms of classification accuracy. In [23], authors define a lightweight tracking algorithm named Kerman (Kernelized Kalman filter), which is a decision tree based hybrid Kernelized Correlation Filter (KCF) algorithm for human object tracking. In [4, 23, 30] several machine learning models are used to detect people during video surveillance activities, showing results in terms of classification achieved.

Finally, in the medical field, gesture recognition is applied to monitor diseases. Authors, in [21], present an innovative approach to preterm infants' limb pose estimation. They

exploit Spatio-temporal information to determine and track limb joint position from depth videos with high reliability. Instead, in [34], the authors illustrate Ultigesture wristband, a hardware/software platform for gesture recognition and remote control. Ultigesture wristband offers full open API for third-party research and application development.

3 Preliminaries

This section presents preliminary notions related to MIDI music, such as the types of MIDI notes and the parameters involved in the structure of MIDI messages. Moreover, we report a discussion regarding *Helmholtz pitch notation* [14], which was used in our system for mapping operations.

3.1 An overview of MIDI

The Musical Instrument Digital Interface (MIDI) is a standard de facto for enabling communication among digital musical instruments and processors of digital music, such as personal computers and sequencers [22]. A MIDI message carries data concerning the peculiarity of a certain sound, such as the vibrato, the tremolo, and so on. However, a MIDI file does not contain any actual waveform generated by the notes of a played musical instrument. Instead, it is a collection of data informing on how a type of sound can be simulated by the digital music processor, which is then responsible for playing the sound by retrieving the representation of the simulated music instrument assigned to those notes from its memory [12].

MIDI is transmitted as asynchronous bytes at 31250 bits per second. One start bit, eight data bits, and one-stop bit result in a maximum transmission rate of 3125 bytes per second. More specifically, a MIDI message is transmitted over 16 different channels in groups of 8 bits, each of which can be of two types called status byte and data byte, respectively. The latter defines the value associated with the message, whereas the former is used to specify the type of message sent. The bytes are distinguished through the first bit, that is, a status byte begins with the bit 1, whereas a data byte begins with the bit 0. A MIDI message is usually composed of a status byte, followed by one or two data bytes, and can belong to one of the two following categories: channel message or system message. Channel messages are sent to single channels and contain information about the musical performance; system messages are aimed at the MIDI system responsible for coordinating the succession of sounds.

Figure 1 shows an example of a MIDI message. In particular, it shows a channel message, composed of one status byte, which is mandatory for every type of message that has been sent and two data byte. The status byte contains information about the operation to be performed and the channel involved with that particular operation, which in this case of Fig. 1 is “play a note on the third channel”. The following two data bytes contain information about the note to be played and the velocity to be applied. The velocity is a value through which it is possible to emulate the amount of force exerted on the key. It can also describe the width of the output or the tone of the sound.

MIDI control changes, also known as or associated with MIDI Controllers or Control Changes, are MIDI messages conveying positional information related to performance control devices such as wheels, sliders, pedals, switches, and other control-oriented devices [29]. This type of information can be used to control a variety of functions, such as vibrato

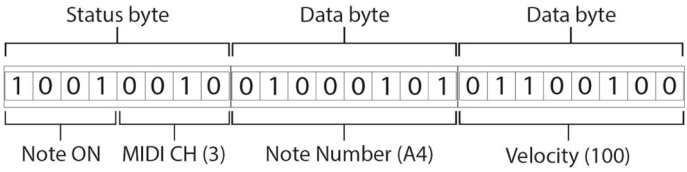

Fig. 1 An example of MIDI message

depth, brightness, and many other parameters. MIDI controller change messages are 128, and each of these has a control number and a control value parameter.

MIDI notes and control change are used in our system to map the spatial coordinates to sounds and to give a different effect to each sound. Details about the proposed system and on how it uses MIDI messages are provided in the next section.

3.2 Scientific pitch notation

In the music field, there exist several musical notations to represent the note in music sheets. One of the most known notation is the *Scientific pitch notation*, which permits to express each note using the literal system. Starting from this, different types of notation have been created, such as the *Tablature* and the *Helmholtz Pitch Notation*. The Scientific pitch notation is one of the most used notations in the multimedia systems, which uses letters to represent the musical notes and numbers to represent octaves. Table 1 shows the name of each note expressed according to the literal system:

The scale in literal system always starts on the note *C* and ends on the *B*, which represent the notes *DO* and *SI*, respectively. Each note can be represented by considering different octaves, using upper-case syntax for low notes, and lower-case syntax for high notes. For example, let us consider note *A*, which represents the note *LA*, it is possible to define different types of *A* by adding "," to represent the notes of the sub-primes or " ' " for primes notes (Fig. 2). Moreover, this notation also supports representation for sharp and flat by adding the suffices "is" and "es", respectively. Notice that, the Scientific pitch notation provides a broad syntax, but in this section, we have only reported the main elements of the notation which will be used in our system.

4 A system for mapping human movements into sounds

We propose a new system named PIANO (maPpIng humAn movements iNto sOunds), which recognizes the movements of people in space and translates them into sounds.

PIANO is a modular system offering standalone modules, designed to be fast and easy to use, relying on non-expensive hardware devices, available on the market. In particular,

Table 1 Literal notation for music notes

Note	DO	RE	MI	FA	SOL	LA	SI
Literal value	C	D	E	F	G	A	B

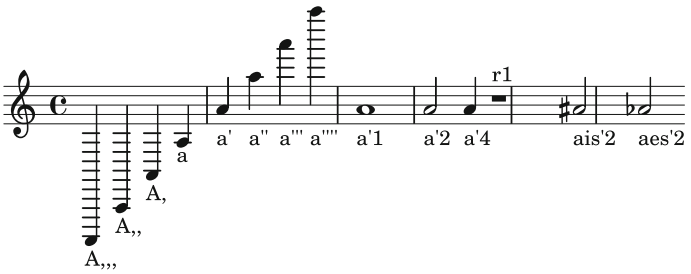

Fig. 2 Example of a music sheet which exploits the *Helmholtz scientific pitch notation*

PIANO consists of three different modules: the first module is a digital tracker relying on a camera to distinguish the participants' bodies and associate them to a virtual figure; the second module maps movements of people into MIDI notes; finally, the third module starts from these to draw music sheets by considering different music clefs. The latter module allows us to concretize the sounds generated by the movements in the space of each people, and to reproduce them at other times. In the next sections, we will discuss the details of each module by analyzing the methodologies and algorithms used by PIANO.

4.1 The object/human detection algorithm

Tracking movements of a person in a live video is not always a simple operation. Thus, several problems must be considered, such as the variability. In fact, a position detection algorithm must be able to trace the object considering the enormous variations in the appearance and position of the objects. Therefore, there are several variables that a tracking algorithm must take into consideration: the point of view, the position, the lighting conditions, the quality or the occlusions of the images. Furthermore, the large number of successive frames that may belong to a live or recorded video can make tracking activities particularly difficult. A tracking system of a moving object must be able not to lose the position of the object in subsequent frames. Thus, an algorithm must be able to consider the variation of all these variables and to perform real-time activities on any environment. These types of operations are even more complicated when needing to identify and/or track groups of objects/people.

Within PIANO we integrated two different types of tracking modules capable of using a large variety of cameras for recognizing people. In fact, PIANO relies on OpenCV,[2] one of the most used libraries for computer vision that can provide a large number of drivers to connect several cameras to our system. The proposed system is able to determine the position of one or many people and to map the position of each person in spatial coordinates. Notice that, both the types of tracking modules exploit existing tracking algorithms. However, we have integrated into this algorithm a strategy to automatically map coordinates into sounds.

Figure 3 shows the phases of the PIANO: detection of people, tracking of these subjects, evaluation of the tracking results to describe events, and drawing of music sheets.

[2] https://opencv.org/

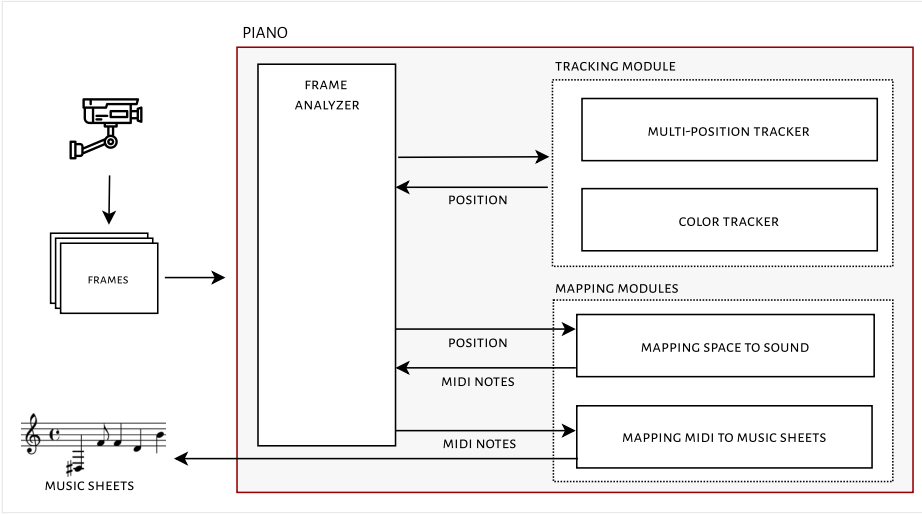

Fig. 3 Architecture of PIANO

In the first phase, the tracking algorithm reads the video as input and converts it to subsequent frames. In particular, the algorithm can work on static video sources or real-time input streams. This guarantees the high adaptability of PIANO to any type of situation and the repetition of itself several times, even after a live video recording. After selecting the input, PIANO allows choosing the tracking mode, i.e. manual or automatic.

Manual tracking enables a full control of the system and the arbitrary choice of whether or not to select a person to track. The user draws the selection rectangle of the objects or people s/he wants to monitor. In particular, each selection is a Region of Interest (ROI) to be tracked. Each selected ROI is associated with a person, and it will be independently recognised by the others. As shown in Fig. 4a, when a ROI is selected, the algorithm creates a virtual figure. Starting from this figure, the algorithm calculates the coordinates and defines the centre of the ROI. This is the reference point of the person's movement. Using different rectangles, it is possible to keep track of many people at the same time, so that PIANO can track separate music for each person. Notice that, the system can use different types of

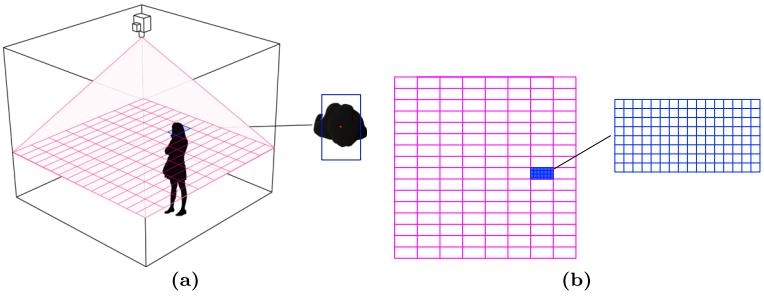

Fig. 4 Virtual grids in the environment

trackers: Boosting [13], Multiple Instance Learning (MIL) [2], Kernelized Correlation Filters (KCF) [15], TLD [18], Median-Flow [17], andMinimum Output Sum of Squared Error (MOSSE) [5].

Table 2 shows the details of each tracker supported by PIANO. Support for different types of trackers enables the algorithm to adapt itself to all tracking situations, regardless of hardware characteristics.

Automatic tracking has been created to track people in a dark room using a camera. Unlike manual tracking, there is no need to select people or objects to track. The proposed tracker receives an input dictionary of colours, and after selecting the input, the individual frames are analysed.

Each frame is resized, blurred, and converted to the HSV (Hue, Saturation, Value) colour space. Next, for each colour defined in the dictionary, the tracker checks objects in each frame. Then, it constructs a mask for the colour from the dictionary, and performs a series of implicit dilations and erosions in order to remove any small blobs left in the mask. Successively, it selects one of the colours in the dictionary and defines the contours of the figure, showing it within the frames. The shapes are defined as circles, and the centre is the reference point for the person's movements. Since this type of tracker has been created for shooting in the dark, two-colour scales have been defined in the algorithm, red and grey.

As shown in Fig. 3, after the tracking phase, it is necessary to evaluate the output of each tracker. The evaluation phase defined in PIANO is described in the next section.

4.2 Mapping space to sound

PIANO uses two different functional systems: digital and analog. The analog system is based on a technology inspired by the "Theremin". It consists of an antenna capable of feeling the proximity of the electric voltage of the human body, and of translating it into an analog audio input, which is pre-amplified by an integrated circuit before issuing the sound through an active speaker. The three objects interacting with each other generate a magnetic field producing a white background noise. Bodies moving through this field alternate this with the magnetic status, and as a consequence, alter the white background noise and sound. In other words, Theremin is exploited for generating background sounds, since it is capable to react to stimuli such as proximity and distance. A more insightful view, provided in Section 5, describes how the analog and digital systems cooperate in allowing users to perceive the space around them. The digital system works on the actual relation of the bodies in the room, associating them to a sound only treated with their position in the room. In particular, we have designed an algorithm to create two virtual layers with which it is possible to track the position of one or more people, and transform it into sounds. This operation is particularly complex so that it is necessary to use a mapping methodology that is fast and precise. To this end, we have developed an approach to quickly map spatial coordinates to sound, and permits to automatically handle the bit-rate of midi notes. After the video stream is read, the proposed methodology creates different virtual layers for each frame: the first layer defines the midi note that must be played, whereas the second layer represents the control change for each note. Notice that, this mapping strategy allows the system to simultaneously monitor the position of people in the virtual layers with the aim of playing both the notes and the effects associated with the movement. Figure 4 shows how the frame is divided. In particular, Fig. 4a shows a representation of a frame from a video stream. We divided each frame into 128 different equal parts, corresponding to the number of existing midi notes. To do this, PIANO automatically calculates the size of each section.

Table 2 Types of trackers available in PIANO

Tracker	Details	Pros	Cons
Boosting [13]	It relies on the same algorithm used from Haar Cascades [33], and like them, the classifier of this tracker is trained online during execution. For each frame, the tracker explores the pixels around the bounding box to evaluate the neighborhood of the region with a score, and in the next frame, it explores the neighborhood with the highest score.	It works well, especially with non-high quality video sources.	This tracker is slow and has been replaced by other advanced trackers based on similar principles available.
Multiple Instance Learning (MIL) [2]	Similar to the Boosting tracker, it considers the current position of the object as a positive example, and a small neighborhood around the current position to generate several potential positive examples.	Performance is pretty good. The tracker is more stable than the Boosting tracker and is able to handle partial occlusion.	Detection error is not reported reliably. The tracker does not recover from complete occlusion.
Kernelized Correlation Filters (KCF) [15]	It is based on MIL tracker. In particular, it uses some mathematical properties to make tracking faster and more accurate at the same time.	Accuracy and speed are both better than MIL and it reports failure better than Boosting and MIL	It does not recover from full occlusion.

Table 2 (continued)

Tracker	Details	Pros	Cons
Tracking, Learning and Detection (TLD) [18]	It divides tracking operation into three different phases: 1) detection phase in which the detector localizes all the objects that have been observed; 2) learning phase in which the detector calculate the errors and updates the tracker to avoid these errors in the future, and 3) tracking phase that allows it to track an object over a larger scale, also when there are occlusions.	It tracks changes in scale and works well when there are occlusions across multiple frames.	It generates a lots of false positives making it almost unusable.
Median-Flow [17]	It tracks the objects in both forward and backward directions in time. It measures the differences between these two directions aiming for minimizing the Forward-BackwardError. The latter allows the tracker to detect tracking failures and select directions in video sequences.	It works very well when the motion is small and there are not occlusions.	It fails if there are large movement.
Minimum Output Sum of Squared Error (MOSSE) [5]	It uses adaptive correlation for object tracking. This produces stable correlation filters during the initialization phase by using a single frame.	It is very fast and works at a higher FPS. Moreover, it quickly detects occlusions, and it is robust to variations in lighting and pose.	It is not as accurate as CSRT or KCF

Formally, let F be the frame extracted from the video, F_{width} and F_{height} the height and the width of F, respectively, both expressed in pixel, S_h and S_w the maximum number of rectangles on F_{width} and F_{height}, i.e. $S_w = 16$ and $S_h = 8$. Then, starting from the centre of the bounding box identified on the frame at the position (H_x, H_y), it is necessary to use Algorithm 1 to define the MIDI note to play. The values related to the user's position inside the frame are normalized with respect to the size of the frame and its subdivision in equal rectangles. Furthermore, the value deriving from the calculation is rounded to the lower integer, thus preventing small changes in position from causing a sudden change of note. More specifically, given the size of each frame expressed in pixel, the number of rectangles in which the space has been divided, and the centre of the bounding box defined in the frame to track one person, the algorithm defines the height and the width of each rectangle in the first grid (lines 1-2), and then calculates the number of the rectangle in which the person moves (lines 3-4). As said above, each rectangle on the grid corresponds to a different note, so that it is possible to define the MIDI note to be played by only considering the coordinates of the rectangle (line 5).

Figure 4b shows the structure of the second layer. In particular, each rectangle of the first layer is in turn divided into 128 rectangles of equal size, which represent the control change. When the centre of the bounding box plays a MIDI note in the first layer, the control change is simultaneously calculated in the second layer.

Algorithm 2 provides the control change value to be played. Similarly to Algorithm 1, it starts by considering the dimensions of each rectangle in the first grid, aiming to define the height and the width of the rectangles in the second grid (lines 1-2). As said above, the control change can take 128 possible values, each defined from a single rectangle in the second grid. Thus, the algorithm calculates the control change to be reproduced by considering the size of the frame and the position of the people (lines 3-5).

Using the Algorithms 1 and 2, we obtain two pairs of coordinates (G_x^1, G_y^1) and (G_x^2, G_y^2), respectively, that allow us to identify the MIDI note and the control change in F to play. It is important to notice that control changes are the messages that allow the MIDI protocol to give expressiveness to its music. In fact, without them, the music resulting from the MIDI messages would appear quite robotic. To this end, the combination of MIDI notes and control change effects allow us to reproduce catchy and fluid music [16].

Algorithm 1 `MAPPING_SPACE_MIDI`.

INPUT: The coordinates H_x, H_y of the person in the space extracted from the bounding box; The height F_{height} and the width F_{width} of the frame; The number of rectangles on the X-axis of the frame S_w; The number of rectangles on Y-axis of the frame S_h
OUTPUT: The MIDI note value to be played

1: $R_{height}^1 \leftarrow \left\lfloor \frac{F_{height}}{S_h} \right\rfloor$
2: $R_{width}^1 \leftarrow \left\lfloor \frac{F_{width}}{S_w} \right\rfloor$
3: $G_x^1 \leftarrow \left\lfloor \frac{H_x}{R_{width}^1} \right\rfloor$
4: $G_y^1 \leftarrow \left\lfloor \frac{H_y}{R_{height}^1} \right\rfloor$
5: $MIDI_{note} \leftarrow S_h * G_x^1 + G_y^1$
6: **return** $MIDI_{note}$

Algorithm 2 MAPPING_SPACE_CONTROLCHANGE.

INPUT: The coordinates H_x, H_y of the person in the space extracted from the bounding box; The number of rectangles on the X-axis of the second virtual layer S_w; The number of rectangles on the Y-axis of the second virtual layer S_h
OUTPUT: The control change value to be reproduced

1: $R^2_{height} \leftarrow \left\lfloor \frac{R^1_{height}}{S_h} \right\rfloor$
2: $R^2_{width} \leftarrow \left\lfloor \frac{R^1_{width}}{S_w} \right\rfloor$
3: $G^2_x \leftarrow \left\lfloor \frac{(H_x - G^1_x \cdot R^1_{width})}{R^2_{width}} \right\rfloor$
4: $G^2_y \leftarrow \left\lfloor \frac{(H_y - G^1_y \cdot R^1_{height})}{R^2_{height}} \right\rfloor$
5: $ControlChange \leftarrow S_h * G^2_x + G^2_y$
6: **return** $ControlChange$

4.3 Mapping MIDI notes to music sheets

The third module allows PIANO to give a concrete form to the movements, aiming to reproduce the generated sounds even after they have been captured. In particular, the module consists of two phases: the first phase aims to map from MIDI notes to Helmholtz's scientific notation (Section 3.2); the second step is to draw a music sheet using the new notation.

As described in Section 3, this type of notation allows PIANO to represent all playable notes with different musical instruments. Since there are instruments that cannot reproduce very high and/or very low notes, it was necessary to select a set of notes that can be played by most instruments. After a consultation with a musician, we have selected 128 different notes with different duration and tonality in order to create playable music sheets (Table 3). In particular. we decided not to get too specific about some types of notes, i.e. sixteenth note, sixty-fourth note, and so on, all of which can be represented by the Helmholtz scientific notation. The reason is due to the need to represent an already wide spectrum of notes in a narrow space such as that of the virtual grid. Including too many short notes would have created too many difficulties in maintaining a certain sound based on the position of the user, rather it would have generated a too sudden succession of notes, which would have led to a perception of the overall sound not very catchy. As described in [27], the operation to create digital melodies starting from MIDI sources it is a quite complex problem. To this end, we have chosen to provide the musicians a reproducible music sheet, leaving them to create the melodies.

This module starts from the numbers of MIDI notes generated from the Algorithm 1, and translates each note according to the criteria defined in Table 3. The new values of each note allow PIANO to draw a music sheet in *LilyPond*[3] language. The latter provides specific syntax for music notes aiming to quickly create multiple music sheets by simultaneously

[3]https://lilypond.org/

Table 3 Set of notes extracted from Helmholtz scientific notation for the mapping operation in PIANO

MIDI	Note	MIDI	Note	MIDI	Note	MIDI	Note	MIDI	Note	MIDI	Note
1	a1	2	a2	3	a4	4	a8	5	b1	6	b2
7	b4	8	b8	9	c1	10	c2	11	c4	12	c8
13	d1	14	d2	15	d4	16	d8	17	e1	18	e2
19	e4	20	e8	21	f1	22	f2	23	f4	24	f8
25	g1	26	g2	27	g4	28	g8	29	a'1	30	a'2
31	a'4	32	a'8	33	b'1	34	b'2	35	b'4	36	b'8
37	c'1	38	c'2	39	c'4	40	c'8	41	d'1	42	d'2
43	d'4	44	d'8	45	e'1	46	e'2	47	e'4	48	e'8
49	f'1	50	f'2	51	f'4	52	f'8	53	g'1	54	g'2
55	g'4	56	g'8	57	ais	58	bis	59	cis	60	dis
61	eis	62	fis	63	gis	64	ais1	65	bis1	66	cis1
67	dis1	68	eis1	69	fis1	70	gis1	71	ais2	72	bis2
73	cis2	74	dis2	75	eis2	76	fis2	77	gis2	78	ais4
79	bis4	80	cis4	81	dis4	82	eis4	83	fis4	84	gis4
85	ais8	86	bis8	87	cis8	88	dis8	89	eis8	90	fis8
91	gis8	92	aes	93	bes	94	ces	95	des	96	ees
97	fes	98	ges	99	aes1	100	bes1	101	ces1	102	des1
103	ees1	104	fes1	105	ges1	106	aes2	107	bes2	108	ces2
109	des2	110	ees2	111	fes2	112	ges2	113	aes4	114	bes4
115	ces4	116	des4	117	ees4	118	fes4	119	ges4	120	aes8
121	bes8	122	ces8	123	des8	124	ees8	125	fes8	126	ges8
127	r1	128	r2								

considering different music clefs. In particular, for each person, PIANO create five music sheets according to five different clefs: *Treble*, *Bass*, *Alto*, *Tenor*, and *GuitarTab*.

Figure 5 shows an example of the tracking process performed by PIANO. In particular, in the example have been considered two different people in the space, distinguished by blue and red colours, respectively. Initially, both people are close and occupy the cells identified with MIDI notes 59 and 60. However, they move to different parts of the space by tracking different trajectories. As they move, the tracking module (Section 4.1) tracks their position, calculates the MIDI note and the control change values for each movement (Algorithms 1 and 2), and draws the notes on the music sheet. Figure 6 shows the final outcomes of this module for the tracks showed in Fig. 5. Notice that, this module allow PIANO to provide a concrete representation of the space related to the movements.

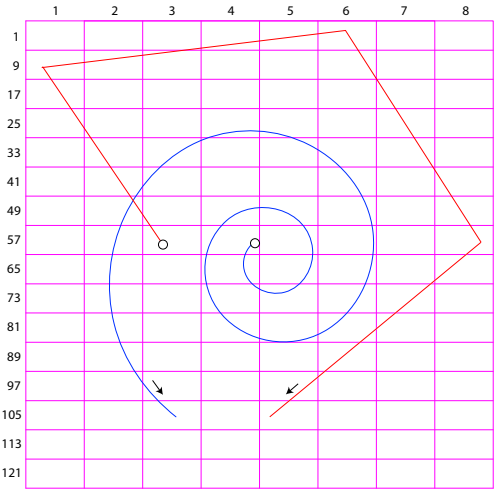

Fig. 5 Example of movements generated by two people

5 Interaction room

In order to present the case study, it is necessary to provide some details about the perception of space aiming to understand how the user perceives space through unconventional instruments such as music.

MIDI	60	69	76	77	70	61
Note	eis	gis1	gis2	ais4	ais2	fis
MIDI	54	53	60	69	76	84
Note	g'4	g'2	eis	gis1	gis2	ais8
MIDI	85	86	78	70	63	54
Note	bis8	cis8	bis4	ais2	ais1	g'4
MIDI	46	38	37	29	28	36
Note	e'4	c'4	c'2	a'2	a'1	c'1
MIDI	35	43	42	50	58	66
Note	b'8	d'8	d'4	f'4	cis	dis1
MIDI	74	82	90	98	99	107
Note	eis2	fis4	gis8	aes1	bes1	ces2

MIDI	59	51	50	42	34	26
Note	dis	f'8	f'4	d'4	b'4	g4
MIDI	25	17	9	10	11	3
Note	g2	e2	c2	c4	c8	a8
MIDI	4	5	6	14	7	23
Note	b1	b2	b4	d4	b8	f8
MIDI	31	39	47	48	56	64
Note	a'8	c'8	e'8	f'1	ais	bis1
MIDI	72	71	79	87	86	94
Note	cis2	bis2	cis4	dis8	cis8	des
MIDI	102	101	109			
Note	ees1	des1	ees2			

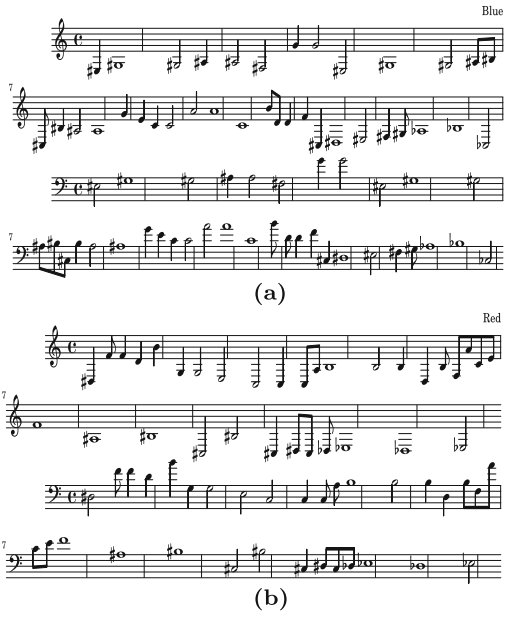

Fig. 6 Music sheets generated according to the movements in Fig. 5

Starting from the definition provided in [31], we consider the space as a room. The term *room* is derived from the archaic English *rum*, which is similar to the German world *Raum* (space). This, in turn, refers to the Latin derivation — *rus* — which can be translated as the act of making space. In this definition, the room is not described as a space associated to a specific role in the context where it is located, like for example, as a component of an apartment. Rather, room here is meant as a volume where phenomena take place. Our attempt is to understand the meaning of movements in the architecture domain and how to design the space based on them. The final outcomes of these models produce a monolithic city, able to grow and to allow for urban colonisation, triggering a wild world attitude to the gentrification of an urban environment. Starting from these ideas, we have been designed a specific environment called *Interaction Room*.

As spatial speculation, the Interaction Room represents a format of feedbacking architecture. It aims to offer an opportunity, a phenomenon, and a sound, to show how it is possible to generate sounds and emotions by using the space and the relative position each one takes within, coherently with the presence of other individuals.

Sounds are generated into two different modes. The original sound is produced digitally through the methodology defined in Section 4.2. However, the produced sound can be distorted analogically. This distortion is achieved thanks to the adoption of above described Theremin. These devices have been integrated inside the objects scattered in the interaction room. The closeness of the user to an object, and therefore to the Theremin, produces an unpredictable distortion to the digitally produced sound. The intention is to merge these two solutions to have more a deep sound that, at the same time, is capable of enhancing the users' perception of the surrounding environment.

The experience aimed to learn how to be in a space together, generating a sound that depends on how people perceive the relation with space. Through artistic and digital approaches, we have given people the opportunity to discuss their space recognition experience.

5.1 Experimental design

The Interaction Room has been designed as cubic structure with dimensions of 3 meters (Fig. 7). Space is interrupted by the presence of three objects, which simulate limitations, edges, and walls, that the participants could use to describe their own layout (Fig. 7b). In particular, these structures are parallelepipedic objects having dimensions of $1.7 \times 0.7 \times 0.3$ meters, and a weighing less than 10 kg.

PIANO has been executed on a computer with an Intel i9-9900k 3.6GHz 8-core CPU with 32 GB of RAM. The system interacted with an infrared camera with a resolution of 3 megapixels and an external sound card to reproduce the sounds. The latter has been connected to four 500-watt audio speakers located in the four corners of the room.

The evaluation session has been performed by involving 30 people of different ages in supervised experience. In particular, people had different qualification: high school diploma (20%), bachelor degree (30%), master degree (40%), and PhD (10%). Moreover, some of the people that have been involved in the experience were musicians, architects, and surveyors with considerable past experience.

The overall experience was based on the production of sounds. The idea was to link the sounds produced by the movements of participants to a specific position in the room, so as

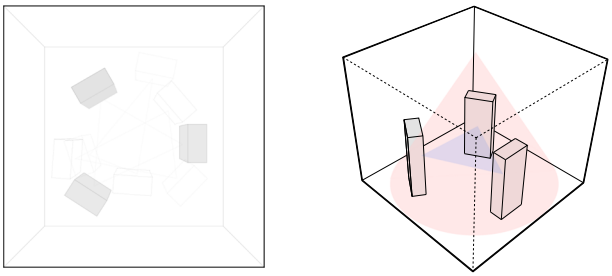

Fig. 7 Interaction room

to relate it to a specific sound frequency (a note). The experience consists of accessing into the room, which appears as dark volume. In this room, people were able to face other participants only as silhouettes, and the obscure environment did not enable them to distinguish proper figures and images. Shadows are generated by a tiny light located on the edges of the room.

The aim of this experiment is to achieve pure perceptive feedbacks from participants, without triggering an "over-automatisation" process in adapting to space. A demonstration video of the interaction room experience is available on YouTube.[4]

5.2 Experimental results

In this section, we present the results of a user study that aims to evaluate how PIANO affects the users' perception of the space through the automatic acquisition of movements and their mapping into music.

The involvement of participants was fundamental to test the usefulness of PIANO since, through their role in participating in the experience, they could possibly suggest solutions maybe initially unseen or ignored.

Before starting the experience, we explained the role of the room and the scope of the experience. Subsequently, the structure of the room was illustrated and the participants were invited to enter the room. In particular, each experience involved a maximum of two participants at a time in order to avoid confusion due to the limited space. After each participant was in the room, we invited them to move freely. Participants were immersed in an unfamiliar environment, and they were forced to re-familiarise themselves with space. In fact, the participants were free to explore the room without any constraints and pressure. At any time, PIANO tracked the movements of each participant, and immediately mapped them into sounds. The sounds were used for moving into the room, in order to become aware of space. The single experience lasted 2 minutes for each participant (Fig. 8).

At the end of the experience, we have provided a specific survey useful for evaluating the effectiveness of PIANO for the perception of the space. In particular, the survey contains basic questions, such as age, gender and qualification, and specific questions concerning the users' perception of sounds and space. In the following, we report the questions used for the evaluation of the user's experience: (1) What are the senses that may affect the human perception of the space? (2) Did you react faster to acoustic than visual stimuli? (3) Do you

[4] https://youtu.be/ZqYjUALQiG8

Fig. 8 An example of multi-position tracker

think this type of experience can help you perceive space? (4) What did you feel during the experience? (5) Did you feel able to control space through sound? (6) What are, in your opinion, the possible uses of this system? For the questions (2), (3), and (5) we adopted single choice answer format, while for (1), (4), and (6), we provided a multiple-choice answer format in order to allow the users to clearly assess their feelings and perceptions during the experience.

Figure 9 shows the user statistics collected about the type of senses used to perceive the space. It is possible to notice that most of the people generally exploit the sight and hearing to become aware of their surroundings.

Figure 10 show the answers about the ability to react on visual and acoustic stimuli. Since we have performed the experiment in a dark environment, a lot of users manifested a faster reaction to the acoustic stimuli. Furthermore, some participants appeared to move away from objects even without the intention to avoid the obstacle we placed. In fact, since the approach to the objects caused distortion on the sound they were producing, participants instinctively moved away from them so as to not spoil the melody they were composing. However, a small part of the participants faster reacted to visual stimuli, since they were probably influenced by the presence of a tiny light in the room. Moreover, this could be due to the fact that people are not comfortable with navigating an environment without using their eyesight, relying exclusively on sounds.

Figure 11 shows the answers about the feelings of the people during the experience. It is possible to notice that most of the participants felt engagement and curiosity. This could be

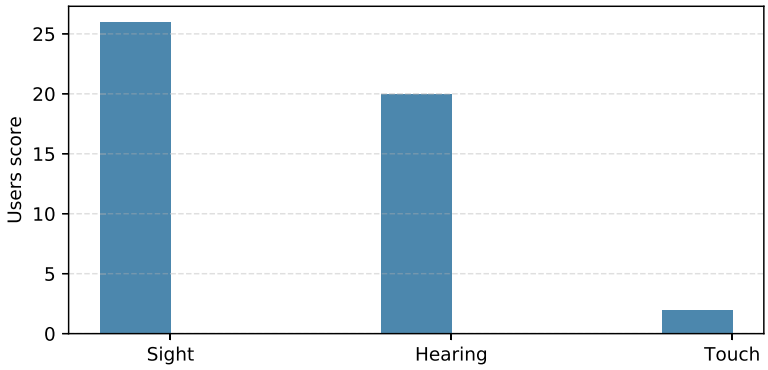

Fig. 9 Question 1: What are the senses that may affect the human perception of the space?

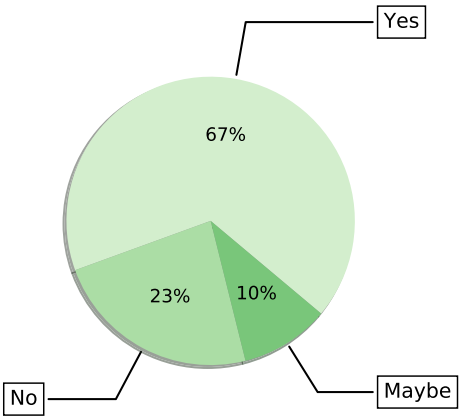

Fig. 10 Question 2: Did you react faster to acoustic than visual stimuli?

due to the fact that most of them considered the usefulness of sounds in perceiving space rather than sight (Figs. 12 and 13).

Figure 14 show the answers about the usage of PIANO in real-life contexts. In particular, we can notice that most of the people consider PIANO useful to support blind people during daily life scenarios. Moreover, another relevant part of the participants also considers useful PIANO to support the architects to develop new types of design styles.

5.3 Discussion

As the most frequent answer, people that took part in this experience were describing the capability of perceiving the space, and interacting with the sound. Describing direct feedback connected with participants' emotions, the room was able to manifest an interactive space, reproducing any movements in the form of sound.

From an architectural and philosophical point of view, the phenomena generated in an environment allow each person to create a personal space. In fact, people who move in an unknown environment are able to learn to move by exploiting the senses to become familiar with the environment, and create their own familiar space. For this reason, it is clear how the environment perceived was not only related to the tectonic of the room itself and its

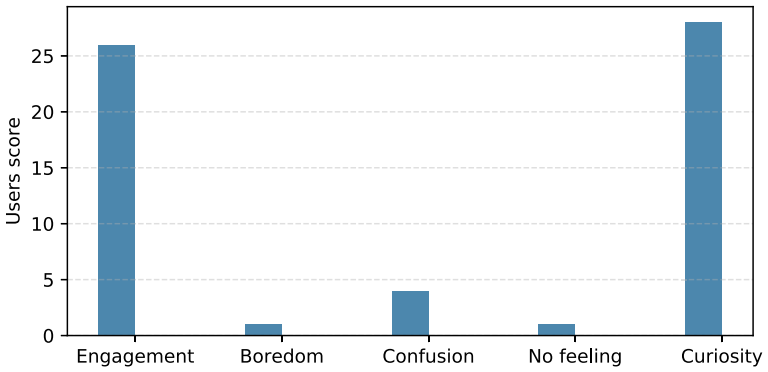

Fig. 11 Question 4: What did you feel during the experience?

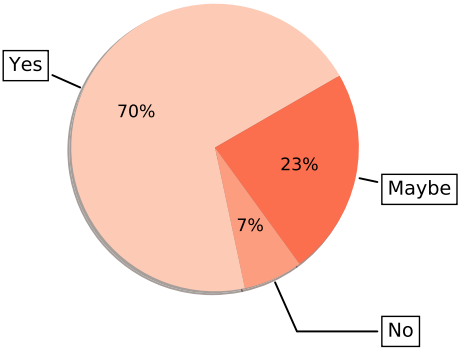

Fig. 12 Question 3: Do you think this type of experience can help you perceive space?

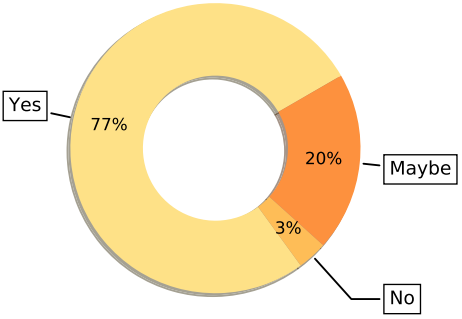

Fig. 13 Question 5: Did you feel able to control space through sound?

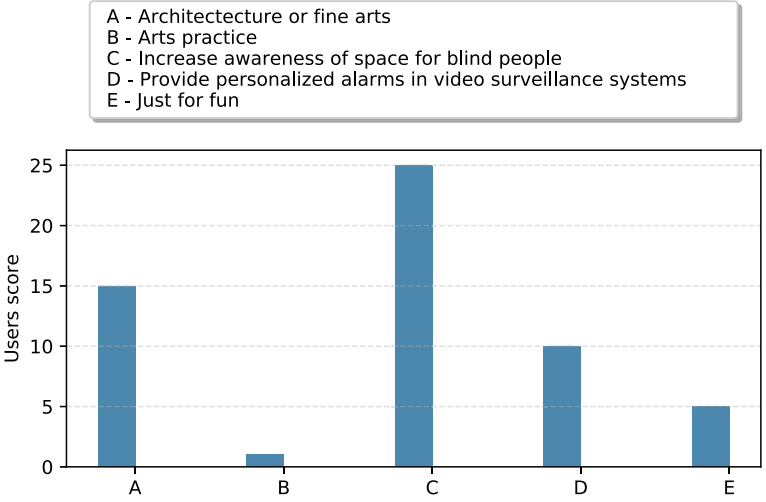

Fig. 14 Question 6: What are, in your opinion, the possible uses of this system?

perception of volume is not associated with predictable habits but only a personal experience strictly linked to the users' perceptions of the space. Thus, it is clear how the act of making space was strictly related to a phenomenon by permitting the adaptation to the environment [31].

Most of the participants claimed to have the impression of perception on the density of the room. This was possible because of the noise produced by the Theremins within the three objects. There was clearly a relation with the quantity of background noise and the quality of the space. Considering again the evaluation from an architectural and philosophical point of view, we can deduce that this imaginary plan allowed participants to perceive themselves in a very different context of the environment. In fact, participants declared that sound gave them a new perspective on their presence within the environment they were moving in, a place where the feedback of their presence was not only ensured by the view of objects moving around them, but also by the variety of sounds that accompanied them during all experience. In other words, this architecture was able to make the participants feel the space around them, and also feeling part of it, without needing to use their eyesight.

6 Conclusion

We proposed an algorithm and a system to recognize human movements in the space, aiming to translate them into sounds. The system has been used in the architectural domain, aiming to provide useful insights for gathering people's perceptions with respect to the surrounding space. The results of our experimental evaluation are obtained from two real experiences carried out in the United Kingdom and Italy. To perform this experience it was necessary to build the interaction room and create every object inside it.

In the future, we would like to perform additional tests in other cities by involving more participants to gather additional information for evaluating the participant's perception associated with the interpretation of sounds. Yet, we would like to investigate the possibility of using different mapping positions to sound for improving users' perception of the space. Moreover, we intend to exploit the proposed system into different application domains, such as assistive technologies for the blind people [32], and for people with neurodegenerative diseases which imply difficulties of motor control, involving in our study the analysis of the behavior of each individual [11]. To this end, we would to define a novel map model in order to enable the possibility to map and distinguish high-level critical scenarios. In fact, an experiment with different mappings of position to sound could help us to select the most effective mapping and to extend the application scenarios for the proposed system.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

References

1. Albacete PL, Chang SK, Polese G (1998) Iconic language design for people with significant speech and multiple impairments. Springer, Berlin, pp 12–32. https://doi.org/10.1007/BFb0055967
2. Babenko B, Yang MH, Belongie S (2011) Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33(8):1619–1632
3. Bagdanov AD, Del Bimbo A, Seidenari L, Usai L (2012) Real-time hand status recognition from rgb-d imagery. In: Proc. of international conference on pattern recognition (ICPR'12), pp 2456–2459
4. Bohush R, Zakharava I (2019) Robust person tracking algorithm based on convolutional neural network for indoor video surveillance systems. In: International conference on pattern recognition and information processing. Springer, pp 289–300
5. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2544–2550
6. Caruccio L, Cirillo S (2020) Incremental discovery of imprecise functional dependencies. Journal of Data and Information Quality (JDIQ) 12(4):1–25
7. Caruccio L, Polese G, Tortora G, Iannone D (2019) Edcar: a knowledge representation framework to enhance automatic video surveillance. Expert Syst Appl 131:190–207
8. Chin-Shyurng F, Lee SE, Wu ML (2019) Real-time musical conducting gesture recognition based on a dynamic time warping classifier using a single-depth camera. Appl Sci 9(3):528
9. Choi I, Bargar R (1995) Interfacing sound synthesis to movement for exploring high-dimensional systems in a virtual environment. In: 1995 IEEE international conference on systems, man and cybernetics. Intelligent systems for the 21st century, vol 3, pp 2772–2777. IEEE
10. Cook T, Roy AR, Welker KM (2019) Music as an emotion regulation strategy: an examination of genres of music and their roles in emotion regulation. Psychol Music 47(1):144–154
11. Costagliola G, Fuccella V, Giordano M, Polese G (2008) Monitoring online tests through data visualization. IEEE Trans Knowl Data Eng 21(6):773–784
12. Flam M (2001) Musical instrument digital interface with speech capability. US Patent 6,191,349
13. Grabner H, Bischof H (2006) On-line boosting and vision. In: Proc. IEEE conference on computer vision and pattern recognition (CVPR'06), pp 260–267. IEEE
14. Helmholtz H (2013) On the sensations of tone. Courier Corporation
15. Henriques J, Caseiro R, Martins P, Batista J (2014) High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3):583–596
16. Huber DM (2007) The MIDI manual: a practical guide to MIDI in the project studio. Taylor & Francis
17. Kalal Z, Mikolajczyk K, Matas J (2010) Forward-backward error: automatic detection of tracking failures. In: Proc. of international conference on pattern recognition, pp 2756–2759
18. Kalal Z, Mikolajczyk K, Matas J (2011) Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(7):1409–1422
19. Marrin T, Picard R (1998) The "conductor's jacket": a device for recording expressive musical gestures. In: Proceedings of international computer music conference
20. Mitra S, Acharya T (2007) Gesture recognition: a survey. IEEE Transactions on Systems Man and Cybernetics C 37(3):311–324. https://doi.org/10.1109/TSMCC.2007.893280
21. Moccia S, Migliorelli L, Carnielli V, Frontoni E (2019) Preterm infants' pose estimation with spatio-temporal features. IEEE Transactions on Biomedical Engineering
22. Moore FR (1988) The dysfunctions of midi. Computer Music Journal 12(1):19–28
23. Nikouei SY, Chen Y, Song S, Faughnan TR (2019) Kerman: a hybrid lightweight tracking algorithm to enable smart surveillance as an edge service. In: 2019 16th IEEE annual consumer communications & networking conference (CCNC), pp 1–6. IEEE
24. Niu W, Long J, Han D, Wang Y (2004) Human activity detection and recognition for video surveillance. In: Proc. of international conference on multimedia and expo, pp 719–722
25. Ortale R, Ritacco E, Pelekis N, Trasarti R, Costa G, Giannotti F, Manco G, Renso C, Theodoridis Y (2008) The daedalus framework: progressive querying and mining of movement data. In: Proc. of international conf. on advances in geographic information systems, pp 1–4
26. Overholt D (2001) The MATRIX: a novel controller for musical expression. In: New interfaces for musical expression, pp 38–41. http://www.nime.org/proceedings/2001/nime2001_038.pdf
27. Ozcan G, Isikhan C, Alpkocak A (2005) Melody extraction on midi music files. In: Seventh IEEE international symposium on multimedia (ISM'05), pp 8–pp. IEEE

28. Raaj Y, Idrees H, Hidalgo G, Sheikh Y (2019) Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In: Proc. of IEEE conference on computer vision and pattern recognition, pp 4620–4628
29. Rumsey F (1994) MIDI systems and control. Butterworth-Heinemann
30. Shehzed A, Jalal A, Kim K (2019) Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection. In: 2019 international conference on applied and engineering mathematics (ICAEM), pp 163–168. IEEE
31. Tattara M, Aureli PV (2017) The room of one's own. The architecture of the (private) room black square
32. Vitiello G, Sebillo M, Fornaro L, Di Gregorio M, Cirillo S, De Rosa M, Fuccella V, Costagliola G (2018) Do you like my outfit? cromnia, a mobile assistant for blind users. In: Proceedings of the 4th EAI international conference on smart objects and technologies for social good, pp 249–254
33. Yildirim ME, Park J, Song J, Yoon B (2014) Gender classification based on binary haar cascade. International Journal of Computer and Communication Engineering 3(2):105
34. Zhao H, Wang S, Zhou G, Zhang D (2019) Ultigesture: a wristband-based platform for continuous gesture control in healthcare. Smart Health 11:45–65