The Phone with the Flow: Combining Touch + Optical Flow in Mobile Instruments

Cagan Arslan, Florent Berthaut, Jean Martinet, Ioan Marius Bilasco, Laurent Grisoni
CRIStAL, CNRS, University of Lille, France
{cagan.arslan,florent.berthaut,jean.martinet,marius.bilasco,laurent.grisoni}@univ-lille.fr

ABSTRACT

Mobile devices have been a promising platform for musical performance thanks to the various sensors readily available on board. In particular, mobile cameras can provide rich input as they can capture a wide variety of user gestures or environment dynamics. However, this raw camera input only provides continuous parameters and requires expensive computation. In this paper, we propose combining camera based motion/gesture input with the touch input, in order to filter movement information both temporally and spatially, thus increasing expressiveness while reducing computation time. We present a design space which demonstrates the diversity of interactions that our technique enables. We also report the results of a user study in which we observe how musicians appropriate the interaction space with an example instrument.

Author Keywords

Mobile instruments, Optical flow, touchscreen

CCS Concepts

•**Applied computing** → **Sound and music computing;**
•**Computing methodologies** → Motion capture;

1. INTRODUCTION

Many sensing capabilities of mobile devices have been explored for musical interaction [3]. Most of the mobile devices today includes various sensors such as a microphone, motion sensors, a touch screen, and multiple cameras. Hence, numerous musical interaction devices were designed around smartphones and tablets [5] [2] [17], in particular using the discrete and continuous input capabilities of the touchscreen. Among these sensors however, the use of built-in camera of mobile devices has been little explored. The simplest approach is to use the camera as a tone-hole. Ananya et al. [2] use the average gray-scale value of the camera input image to detect if the lens is covered. Covering or uncovering of the lens modifies the pitch of the sound similarly to a tone-hole. Keefe and Essl [9] used low-level visual features of the input image such as the edginess to create mobile music performances with visual contributions. Camera image has also been used to track visual reference points to gather position and motion information that are mapped to MIDI controls [15] [14].

However, to our knowledge, built-in cameras of mobile devices have not been used for sonification of moving elements in the camera image. Controlling sound through movement has always been of interest and can be traced back to Kurenniemi's DIMI-O [12] and Rokeby's Very Nervous System [1]. More recent examples such as [13][7][8] [4] use difference between images and optical flow to represent the moving parts of the image, but the methods have been either too simple to extract rich information or too heavy to be run on mobile devices. The details of optical flow methods will be further discussed in section 2.2.

In this paper, we propose combining visual movement detection with the touchscreen of mobile devices in order to reduce the computation time and to open expressive opportunities. We analyze these possibilities using a design space. We also study how users appropriate the instrument for two different movement sources: the user's gestures and the environment.

2. THE PHONE WITH THE FLOW

In this section, we first present the main approach behind our system. It is currently implemented as an Android App (`www.caganarslan.info/pwf.html`). Optical flow features, extracted as described in the next section, are then mapped to sound synthesis parameters. The synthesis can be done either directly on the mobile device, with restrictions on the complexity of the synthesis due to limited computing capabilities, or the features can be sent to external musical software via OpenSoundControl messages.

We also describe a design space of the interaction and mappings opportunities afforded by our system.

2.1 From the Scene to the Touchscreen

Our system relies on real-time spatial and temporal filtering of rich motion information provided by the built-in cameras of mobile devices. Fig. 1 depicts an example scenario of use of *The Phone with the Flow*.

Movement-rich Environment: The physical scene captured by the built-in camera offers a large range of movements, with various periodicity, directions and sources. The sources can be artificial (displays, mechanisms), natural (weather, animals, plants), or originate from people (user's body, other musicians, spectators, by-standers).

Choosing the Interaction Space Mobile cameras have a limited field of view but their portability enables exploration of the surroundings by a point-and-shoot approach. Unlike fixed installations, the mobile camera's field of view can be changed without effort by simply changing its position and orientation. When the camera aims at a part of the movement-rich environment, objects in the field of view are captured, providing the user with a visual feedback of

NIME'18, June 3-6, 2018, Blacksburg, Virginia, USA.

(a) A person, a computer screen, a fan, objects passing by the window are source movement sources.

(b) The user chooses a movement

(c) The user selects a region on the screen

Figure 1: From scene to the the touchscreen.

their movement sources. The user is then free to interact in the combined interaction volume [11] of the camera and the touchscreen.

Filtering the Movements Once the movement sources are displayed on the touchscreen, the user focuses on a region of interest by touching it. The touch selects a region of the image to be analyzed further for detailed movement information. The user can change the position of the region by dragging their finger on the screen, alter its size, combine multiple regions and switch between them. The touch input thus enables filtering of the movements.

2.2 Optical Flow

Optical flow is the distribution of apparent velocities of moving brightness patterns in an image [6]. The movement indicates a motion resulting from the relation between captured objects and the viewer. The motion at a point can be represented with Δx and Δy, the horizontal and the vertical displacement respectively. Therefore displacement vector can be expressed as $v(\Delta x, \Delta y)$, but also in polar coordinates $v(r, \Theta)$ to represent the amplitude and the direction of the movement for a pixel. Fig. 2 shows a color coded representation of the optical flow.

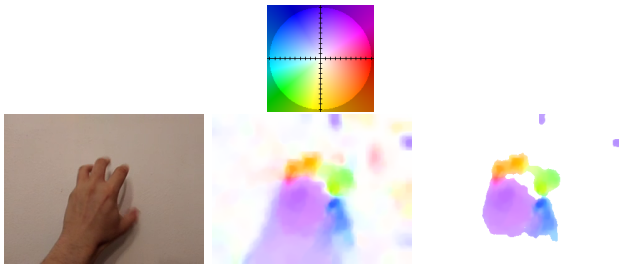

Figure 2: Color representation of the optical flow. (Top) Color space: hue indicates direction, saturation indicates amplitude. (Bottom) Grasp gesture, estimated flow, filtered flow

Optical flow has been used in various musical interfaces in the past. ParticleTecture [7] uses Horn-Schunck's optical flow method to determine the apparent movement in the image. The flow output is interpreted as a Game of Life, where living cells correspond to pixels in motion. The cells are tied to sonic grains that emit sound when activated. However, the only direction information they use is the horizontal component, which is used to trigger grains into proliferation or stasis. Thus, while their work uses optical flow for granular synthesis, they did not fully take advantage directional richness of flow properties. In Sonified Motion Fields[13] Pelletier uses FAST corner detector [16] to extract feature points and estimates the optical flow by using a pyramidal Lucas-Kanade method[18] to track the sparsely identified points. He also discusses the potential mappings of the flow field to sound parameters. However he concludes that the flow fields' temporal resolution is poor and the feature detection is not robust. $CaMuS^2$[14] estimates the optical flow from a camera phone. A 176x144 pixels image is divided into 22x18 blocks and cross-correlations between successive pairs of block images are calculated to obtain the optical flow. Because only 4 points are sampled in each block, the algorithm performs quickly. The simplicity of the method allows obtaining the global translation and rotation at 15fps, but the system is unable to provide rich motion information from the moving objects in the image.

In recent years, there have been advances both in smartphone hardware (improved cameras, CPU and GPU) and optical flow algorithms, which open up new possibilities for movement detection in real-time.

In 2016, Kroeger et al. [10] introduced DIS-Flow, an optical flow estimation method that is one order of magnitude faster than others and produces roughly similar flow quality. Their objective is to trade-off a less accurate flow estimation for large decreases in run-time for time critical tasks. This method provides a dense flow output, in which every pixel is associated to a displacement vector $v(\Delta x, \Delta y)$. This advancement enables us to envision the creation of an innovative design space that allows richer interaction and sonification methods.

Optical flow computation is noisy and presents important motion discontinuities as it measures the displacement in the 2D image plane of the real 3D world. Filtering out noise and focusing on the coherent motion information is the key to successful optical flow interaction. Our approach to extracting features from raw flow data is as follows. First of all, for every region of interest, the pixels that have a smaller displacement value than a threshold are discarded to eliminate the noise (2). We identified key information to gather from the optical flow output. We collect three global values: the amount of moving pixels, the average direction and the average magnitude. The amount of moving pixels corresponds to the ratio of the number of remaining pixels to the image area. The average direction and magnitude are the direction and amplitude of average displacement of the moving pixels. In order to obtain the distribution of the movement within the region of interest, we construct a normalized 8-bin histogram of directions. Adding the three global values to the histogram information, a feature vector of eleven elements is obtained for various mappings.

2.3 Design Space

We present a four dimensional design space to explore the musical expression possibilities presented by *Phone with the Flow*. The dimensions can be combined to provide a large

set of interaction techniques.

2.3.1 Movement Source

A smartphone camera enables the user to freely change the subject of the motion. We distinguish two values for this dimension:

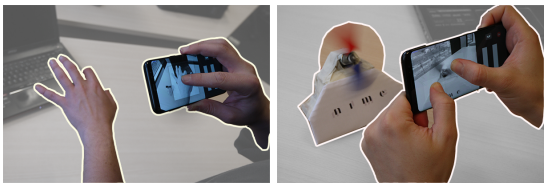

Figure 3: Movement sources. (Left) Self-motion, (Right) World-motion.

Self-motion: The holder of the camera is also the creator of the movements. The user can use its body parts to perform dynamic gestures. In the case where the user holds the camera with one hand, the other hand is free to create optical flow. The fingers permit creating movement in multiple directions simultaneously. Feet, legs, head and even the torso may also be the source of the movement. It is also possible to perform gestures by moving the device, as its relative motion with respect to the environment also creates an optical flow.

World-motion: The camera aims at an external source of motion. These sources can act intentionally to create controlled movements, such as a dancer, or be unaware of how they contribute to the process, such as spectators and other musicians on the stage. Objects that are continuously in motion may also be used to create sound, for instance, a busy highway, a waterfall, an assembly line and so on. Additionally, visual sources such as projections and displays may serve as sources. In a rich environment where there are multiple sources, the mobility of the device allows transition between them.

2.3.2 Camera Movement

This dimension is about how we adapt camera movement to complement the movement search.

Fixed: The camera is held at a fixed position. All of the optical flow is created by the moving objects in the static field of view.

Independent: The camera moves independently from the content of the environment. The optical flow is created by the motion of the camera with respect to the environment. Horizontal and vertical movements of the camera produce an output similar to that of an accelerometer, but the movement in the depth results in an optical flow in every direction as it produces a zoom-in-zoom-out effect. It can also be rotated around three axis.

Adaptive: The camera can track a moving object to strip it from either its translation or rotation. Or, if the object in motion is not rigid, its deformation can be stripped by adapting the camera movement to its trajectory. However, if the object does not fill the camera's field of view, the independent motion mentioned above distorts the optical flow. Hence, adaptive movement requires precise control.

2.3.3 Touch

The touchscreen offers valuable possibilities for filtering the image to discard the movements that are not desirable for the performer. The touch dimension has three values.

None: The touchscreen is not used along with the camera. All the movements in the camera image are processed continuously. This is ideal in controlled environments where no intervention is required.

Activation: The touchscreen is used for temporal-only selection of movement. All movements captured in the camera image are therefore processed, but only while there is a finger touching the screen. This allows for discrete activation/deactivation of otherwise mostly continuous changes in the captured movements.

Isolation: Fingers are used to select a region of the camera image to focus on a movement source and discard the rest of the movements in the view. When there is no contact between the finger and the screen, nothing is processed. This enables both spatial and temporal filtering, making the system more robust to noise. Another advantage of isolation is decreasing the number of pixels to process in order to lower the latency. Multiple regions may be activated at the same time.

2.3.4 Mapping

The mapping dimension describes the mappings between the finger position on the screen and sound parameters.

Independent from spatial position: Sound parameters are influenced only by the optical flow. Regions that encounter the same movement sound the same independently from their position.

Linked to spatial position: X and Y axes of the touchscreen are mapped to the sound parameters. The position of the finger determines the values along these axes. Consequently, the same movement results in different sounds in different regions of the camera image.

3. EVALUATION

We evaluate what techniques are used by musicians when exploring the instrument. The goal of the evaluation is to determine how the users utilize the interaction space created by the combination of the camera and the touch screen.

10 participants (2 left-handed) volunteered to take part in the experiment. Their ages varied between 23-52 (mean 35, sd 9). All participants were involved in musical activities and were regular users of smartphones. The experiment was conducted on a 5.8" Galaxy Edge S8 smartphone running Android. The built-in camera provided images of 320x240 pixels at 30 fps. The feature vectors were sent via OSC messages to a standard PC running our Pure Data patches.

3.1 Protocol

We presented a mobile instrument which demonstrated our design space to the participants. In this instrument, the sound is produced by granular synthesis with 3 voices (one per finger). The mapping is *linked to spatial position* such that the horizontal and vertical touch position on the screen respectively changes the initial position in a sample and the initial frequency of the grains. While the finger is touching, the grains that make up each voice then move around the sound (in position and frequency, as if browsing a spectrogram) according to the optical flow. The flow histogram gives the amplitude of displacement in 8 directions for the grains. The average flow amplitude changes the duration of the grains and the amount of moving pixels controls the volume. Participants were also able to change the size of the region that is activated by a touch.

The experiment was conducted in two stages. In the first, the participants were provided with a display that could generate animations and a small fan as world movement sources. In the second they were instructed to use their bodies to create movement. For each stage they were given 5-10 minutes to explore and to prepare a 5 minute performance. During the performance, all parameter variations

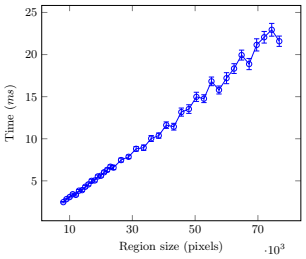

Figure 4: Computation time of the features vs region size

Questions	Mean	SD
I made a lot of physical effort	2.65	1.05
I needed to concentrate	5.25	0.60
I felt frustrated/annoyed	2.6	0.99
I enjoyed interacting with the instrument	5.75	0.84
I was able to create a wide variety of sounds	4.35	1.18
I found my own techniques to play	4.9	1.14

Table 1: Likert Scale results

were logged in the Pure Data patch. The sound samples were altered between the steps. Both the order of steps and the samples were counterbalanced across participants. After the open exploration and the performances, participants responded to 7-point Likert-scale statements for which the results can be seen in table 1.

3.2 Results and Discussion

Without any instruction, all of the participants used their dominant hand to hold the smartphone and to use the touch input. They stated that orienting the smartphone and using the touchscreen at the same time was a difficult task for the non-dominant hand.

We measured the frequency of simultaneous fingers used on the screen. While performing hand gestures in the self-motion condition, the participants mostly used only one finger (55% 1 finger, 32% 2 fingers, 13% 3 fingers) to isolate the regions. However, when participants moved the camera with two hands, they were able to use multiple fingers. In the world-motion, they often used both thumbs (42% 1 finger, 48% 2 fingers, 10% 3 fingers).

Figure 4 shows the average time it takes to calculate optical flow and to produce the feature vector for regions of different size. We calculated the average region size used by the participants to be 16800 (140x120) pixels. This shows that using image regions instead of the entire image (76800 pixels) significantly decreases the computational time from 23ms to 5ms.

70% of the participants preferred using world objects instead of gestures. While the participants who preferred world objects argued that the instrument required concentration when performing hand gestures and using the touch input at the same time. The opponents all stated that gestural interaction allowed a finer control on the created movement.

4. CONCLUSIONS

In this paper, we presented *The Phone With the Flow*, a novel approach to sonify visual movements by combining the touchscreen and the built-in cameras of mobile devices. In this context, the touchscreen proves to be a valuable tool to filter the visual input temporally and spatially. We also discussed the different interaction techniques enabled by our design space. An interesting future work would be to combine the use of optical flow, i.e. motion features, with the other features in the images captured by the built-in camera, bringing objects textures, colors and shapes to the sonification process.

5. REFERENCES

[1] Very nervous system. `http://www.davidrokeby.com/vns.html`. Accessed: 2018-3-28.

[2] M. Ananya, G. Essl, and M. Rohs. Microphone as sensor in mobile phone performance. In *Proceedings of NIME*, pages 185–188, Genoa, Italy, 2008.

[3] G. Essl and M. Rohs. Interactivity for mobile music-making. *Organised Sound*, 14(2):197–207, 2009.

[4] M. Funk, K. Kuwabara, and M. J. Lyons. Sonification of facial actions for musical expression. In *Proceedings of NIME*, pages 127–131, Vancouver, BC, Canada, 2005.

[5] G. Geiger. Using the touch screen as a controller for portable computer music instruments. In *Proceedings of NIME*, pages 61–64, Paris, France, 2006.

[6] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185 – 203, 1981.

[7] J. Jakovich and K. Beilharz. Particletecture : Interactive granular soundspaces for architectural design. In *Proceedings of NIME*, pages 185–190, New York City, NY, United States, 2007.

[8] A. R. Jensenius. Motion-sound interaction using sonification based on motiongrams. In *Proceedings of ACHI*, pages 170–175. IARIA, 2012.

[9] P. O. Keefe and G. Essl. The visual in mobile music performance. In *Proceedings of NIME*, pages 191–196, Oslo, Norway, 2011.

[10] T. Kroeger, R. Timofte, D. Dai, and L. V. Gool. Fast optical flow using dense inverse search. In *Proceedings of ECCV*, 2016.

[11] N. Marquardt, R. Jota, S. Greenberg, and J. A. Jorge. The continuous interaction space: Interaction techniques unifying touch and gesture on and above a digital surface. INTERACT'11, pages 461–476, Berlin, Heidelberg, 2011.

[12] M. Ojanen, J. Suominen, T. Kallio, and K. Lassfolk. Design principles and user interfaces of erkki kurenniemi's electronic musical instruments of the 1960's and 1970's. In *Proceedings of NIME*, pages 88–93, New York, NY, USA, 2007. ACM.

[13] J.-M. Pelletier. Sonified motion flow fields as a means of musical expression. In *NIME*, pages 158–163, 2008.

[14] M. Rohs and G. Essl. Camus 2 – optical flow and collaboration in camera phone music performance. In *Proceedings of NIME*, pages 160–163, New York City, NY, United States, 2007.

[15] M. Rohs, G. Essl, and M. Roth. Camus: Live music performance using camera phones and visual grid tracking. In *NIME*, pages 31–36, Paris, France, 2006.

[16] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Computer Vision – ECCV 2006*, pages 430–443, Berlin, Heidelberg, 2006.

[17] G. Wang. Ocarina: Designing the iphone's magic flute. *Computer Music Journal*, 38(2):8–21, 2014.

[18] J. Yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.