Seeing the Sound: a New Multimodal Imaging Device for Computer Vision

Andrea Zunino[1,2] andrea.zunino@iit.it
Marco Crocco [1] marco.crocco@iit.it
Samuele Martelli[1] samuele.martelli@iit.it
Andrea Trucco[1,2] andrea.trucco@iit.it
Alessio Del Bue[1] alessio.delbue@iit.it
Vittorio Murino[1,3] vittorio.murino@iit.it

[1] Pattern Analysis & Computer Vision - Istituto Italiano di Tecnologia
Via Morego 30, 16163, Genova, Italy
[2] Dipartimento DITEN - Università degli Studi di Genova
Via all'Opera Pia 11A, 16145, Genova, Italy
[3] Dipartimento di Informatica - Università degli Studi di Verona
Strada le Grazie 15, 37134, Verona, Italy

Abstract

Audio imaging can play a fundamental role in computer vision, in particular in automated surveillance, boosting the accuracy of current systems based on standard optical cameras. We present here a new hybrid device for acoustic-optic imaging, whose characteristics are tailored to automated surveillance. In particular, the device allows real-time, high frame rate generation of an acoustic map, overlaid over a standard optical image using a geometric calibration of audio and video streams. We demonstrate the potentialities of the device for target tracking on three challenging setup showing the advantages of using acoustic images against baseline algorithms on image tracking. In particular, the proposed approach is able to overcome, often dramatically, visual tracking with state-of-art algorithms, dealing efficiently with occlusions, abrupt variations in visual appearence and camouflage. These results pave the way to a widespread use of acoustic imaging in application scenarios such as in surveillance and security.

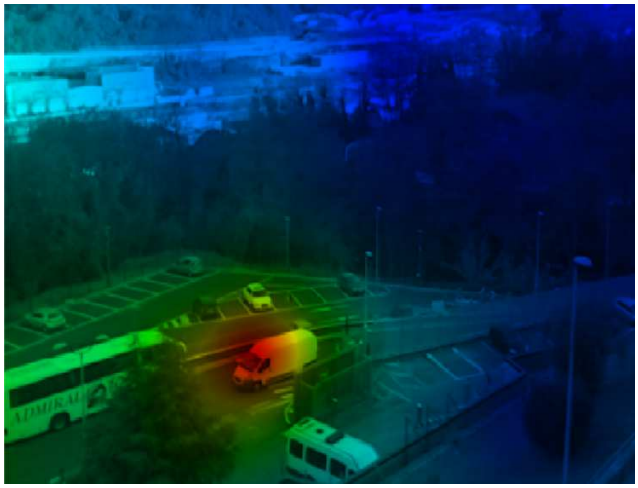

Figure 1. Example of an acoustic map acquired in an outdoor environment overlapped with an optical image. The color-coded map provides at each pixel the sound intensity of the emitting sources. Notice that the red peak is localized at the engine of the van where the emitting source is located.

1. Introduction

Current systems for computer vision, in particular the ones devoted to automated surveillance and security, are based on a network of sensors, typically including cameras (standard optical, thermal and infra-red) together with other devices such as ultra-sound barriers or even radars, whose data streams are typically acquired and processed from a central unit.

Among these sensors modalities, acoustic imaging has surprisingly received scarce attention in the surveillance community, despite the several advantages it bears. These images, resulting from acoustic beamforming applied to the signals acquired by a set of microphones, encode at each pixel the sound intensity coming from each spatial direction (check Fig. 1 for an example). Acoustic imaging may increase the robustness and reliability toward adverse weather conditions because sound propagation is not or barely affected by fog, dust and snow. Moreover, it guarantees 24h

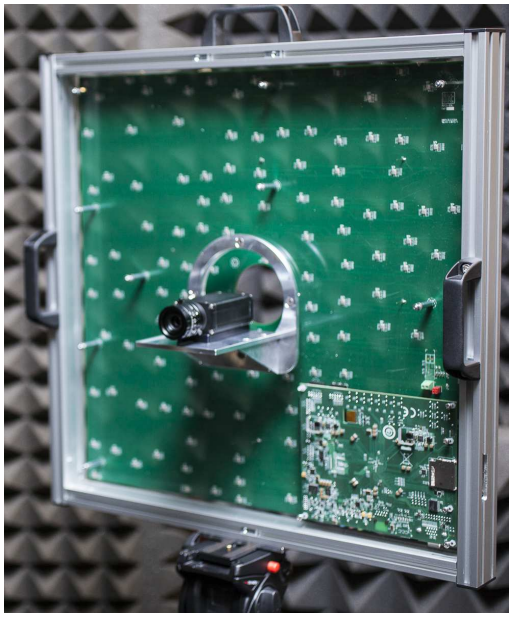

Figure 2. Prototype of the acoustic-optic imaging device. Notice the aperiodic microphone layout, the on board processor in the lower right corner and the optical camera at the center of the device (the camera is placed in the opposite direction for ease of visualization).

functioning because sound is not affected by variations in natural or artificial illumination. Acoustic imaging is also robust to camouflage because a sound source can be localized even if it is hidden behind other objects and it can detect events with little or no visual counterpart, e.g. a gunshot in a crowded scene. Furthermore, it is a completely passive technology, differently from active radars, optical and infrared cameras which require light emitters. For this reason a passive acoustic device does not allow identification from hostile third parties if the sensor is not visible.

Such features make acoustic imaging devices suited to extend the functionalities of current surveillance systems in those scenarios where the target of interest is characterized also by a sound signature. In particular, the coupling together with a video camera may enable compelling applications such as the visual localisation of events that are difficult to understand only using the video signal (e.g. a gunshot in a crowd). Moreover event detection, especially for highly threatening events such as vandalism and riots, have a much more distinguishable audio signature than can help discrimination if coupled with video. However these advantages have to cope with technological limitations of the devices that generate the acoustic map. Commercially available acoustic imaging devices, known as acoustic cameras [1, 2], are mainly intended for industrial testing or environmental noise measurement, performed for small periods of time with temporary installations and requiring a human operator on site. They need a PC connection (no stand-alone functioning) and usually perform off-line processing of data. Moreover they come at high costs since their use is limited to niche markets and as a consequence they can not gain from economies of scale.

Here instead we present a new acoustic-optic device that overcomes the above hardware/cost limitations[1] and we show how this sensor can provide remarkable improvements in a standard Computer Vision tasks such as image tracking[2]. The proposed device produces a real time and high frame rate stream of acoustic images geometrically overlapped, by design, with the optical ones. The optimized microphone layout and processing parameters necessary for beamforming allow to obtain an optimal acoustic image quality in terms of spatial resolution, dynamic range and robustness to diverse environmental conditions, while keeping limited the amount of hardware and software resources. Hence all the data processing is performed on board, by making it a true stand-alone device (check Fig. 2).

Among the several tasks that can be tackled using acoustic images, we have chosen visual tracking since improvements are largely noticeable against the methods using visual data only. Despite notable improvements and evaluations over challenging datasets [20, 15], visual tracking remains in many situations a challenging task, mainly due to the abrupt changes in appearance of the object of interest or occlusions. In many cases even State-of-Art approaches like [11, 13, 21, 3, 9] dramatically fail after few frames. Differently, many targets like vehicles, drones, humans speaking have a very stable sound emission that makes audio tracking much simpler than the visual one. Moreover occlusions do not impair totally the sound propagation, allowing also tracking of hidden objects. Finally the spectral signature of targets can be exploited to distinguish them from other potentially interfering audio sources.

A number of audio tracking methods has been proposed in recent years, grounded on localization by beamforming methods and/or Time Difference of Arrival (TDOAs) estimation. Several issues like multi target tracking [7, 14], robustness to reverberation [17], ground noise and interfering sources [19], time varying and intermittent sources [12] have been addressed, making audio tracking a mature research field. However, the great part of systems for audio tracking is based on a limited number of microphones (typically < 10), typically spread in the environment. Such kind of layout limits the accuracy of localization and the number of sources that can be jointly localized. Finally, the relation between actual position of the target and measurements from the microphones is characterized by strong

[1] The hardware manufacturing per item is roughly 1000 $ considering a production of about 100 items.

[2] Notice that acoustic images can also be used for other tasks such as detection, localization, background subtraction.

non-linearities that forces to use ad hoc tracking algorithms with complex models of measurement noise [4].

Differently, we propose a compact device, easily deployable, producing an *acoustic image* geometrically overlapped by design with the optical one. This fact has multiple advantages:

1. Audio and video information can be easily fused together thanks to the pixel-to-pixel correspondence.
2. Tracking can be performed taking as measurement the cues of the acoustic image instead of cues of the single microphones' raw signal. This allows a transfer knowledge from the much more developed field of video tracking.
3. The relation between measurement and state in tracking is straightforward: actually, in absence of noise they are identical if tracking is performed *onto* the image plane.
4. The overlap between acoustic and optical image allows a fair comparison between audio and video trackers. In particular, we demonstrate here that audio imaging outperforms visual tracking on three different scenarios.

The paper is organized as follows. Sec. 2 describes the new device and the solutions adopted for the acoustic-optic image generation. Sec. 3 presents the audio tracking approach and the three tracking methods considered for evaluation. Sec. 4 shows the results on audio and video tracking. Finally Sec. 5 draws conclusions and directions for future work.

2. Acoustic camera and beamforming

2.1. Beamforming and acoustic image generation

An acoustic map is built up by a spatial-temporal filtering procedure, known as beamforming [16], that takes as input the signals acquired by an array of microphones. In detail, consider a planar array of L microphones, placed in the (x, y) plane with the array center at coordinates $(0, 0, 0)$. Microphones are omnidirectional and placed at coordinates $\mathbf{x}_l^{mic} = (x_l^{mic}, y_l^{mic}, 0)$ with $l = 1 \cdots L$. According to the well known Filter-and-Sum beamforming [16], an acoustic map $M_a(m, n)$ can be expressed as follows:

$$M_a(m,n) = \int_{t \in T} \left| \sum_{l=1}^{L} w_l(t) * s_l(t - \tau_l(m,n)) \right|^2 dt, \quad (1)$$

where $s_l(t)$ is the signal received at microphone l, $w_l(t)$ is the finite impulse response of the FIR filter l in cascade to the l-th microphone, $*$ denotes the convolution operation, c is the sound velocity in air, T is the current time window and $\tau_l(m, n)$ is the delay imposed to the signal from the l-th microphone, related to pixel (m, n) of the acoustic map. In practice, the set of delays $\tau_l(m, n)$ for $l = 1, \cdots, L$ aligns the signals coming from a given point $\mathbf{x}_{m,n}$ in the 3D space so that they can be summed coherently in the beamforming procedure and their energy, calculated by integration over the time window T, be visualized at pixel (m, n). In particular:

$$\tau_l(m,n) = \left(\left\| \mathbf{x}_l^{mic} - \mathbf{x}_{m,n} \right\| - \left\| \mathbf{x}_{m,n} \right\| \right) / c. \quad (2)$$

All the other signals coming from points different from $\mathbf{x}_{m,n}$ will not sum coherently and their contribution at the (m, n)-th pixel intensity will be, in an ideal case, negligible. In a real case, due to the finite number of microphones involved and the finite aperture of the device, the acoustic imaging method will have a given Point Spread Function (PSF), that spreads the contribution of each source over all the image. The PSF affects the spatial resolution of the acoustic image, as well as its dynamic range, the latter accounting for the possibility to visualize several sources of different power in the same time frame. The PSF can be shaped tuning both the FIR filter impulse responses $w_l(t)$ and the microphone layout. We adopted a data independent beamforming techniques in which the set of $w_l(t)$ is fixed and not dependent on the data statistics. Such choice allows a more stable performance of the system and a notable computational saving. In particular, we jointly optimized filters coefficients and microphone positions following the method of [6], which guarantees an optimal trade off between spatial resolution and dynamic range of the acoustic image while keeping the number of microphones reasonable. In addition it assures the robustness of the solution towards deviations of the microphone parameters from their theoretical values. The latter property is of paramount importance when using low cost microphones, as done with the present device.

Setting appropriately the points $\mathbf{x}_{m,n}$ we can fix the projective relation between points in the 3D space and points in the acoustic image. If we fix a focusing distance z_{foc} for the acoustic image and set the points $\mathbf{x}_{m,n}$ as follows:

$$\mathbf{x}_{m,n} = (m z_{foc}, n z_{foc}, z_{foc}), \quad (3)$$

we obtain a relation that is equivalent to the one of an ideal projective camera for points having $z = z_{foc}$. In fact placing the camera center in $(0, 0, 0)$ and assuming a focal distance of the image plane of 1 we have that a generic 3D point (x, y, z) is projected in $(x/z, y/z)$ in the image plane. Hence, the 3D point $(m z_{foc}, n z_{foc}, z_{foc})$ is projected in the same pixel (m, n) for both the acoustic and optic image, i.e. acoustic and optic images are mutually calibrated by design, without the need of further processing. For a source placed at $z \neq z_{foc}$, the point spread function of the acoustic image will broaden, analogously to an out-of-focus optical image,

but its peak will remain centered around the correct pixel (m, n).
The time implementation of the beamforming procedure in Eq. 1 is computationally demanding, due both to the convolution operation and to the need of an high sampling rate, far above the Nyquist rate, in order to evaluate the delayed signals $s_l(t - \tau_l(m, n))$ with a reasonable approximation. Moreover if we want to change the frequency band over which the acoustic image is calculated, the whole beamforming operation must be replicated, or, in alternative, each of the (m, n) signals must be filtered before squared modulus integration in Eq. 1, both the two options being computationally demanding. For this reason we adopted a frequency implementation of the filter-and-sum beamforming:

$$M_a(m,n) = \int_{f \in F} \left| \sum_{l=1}^{L} W_l(f) S_l(f) e^{-j2\pi f \tau_l(m,n)} \right|^2 df, \tag{4}$$

where $S_l(f)$ and $W_l(f)$ are the Discrete Fourier Transform (DFT) of $s_l(t)$ and $w_l(t)$ respectively and F is the desired frequency band. Notice that the delays have been replaced by complex exponentials, thus avoiding oversampling and approximation errors. In addition the frequency band can be easily tuned, simply changing the domain of integration in Eq. 4.

2.2. The Acoustic-Optic camera

Based on the beaforming implementation and microphone layout optimization described in the previous section, we designed and assembled a prototype of a stand-alone acoustic-optic camera.

The developed device is composed of three main modules: a planar array of 0.45×0.45 meters composed of 128 MEMS low-cost digital microphones displaced according to an optimized aperiodic layout, a video camera placed at the device center and an on-board hybrid embedded processor as illustrated in Fig. 2. In order to deploy the sensor in outdoor environments, we have also developed an ad-hoc housing that allows an easy deployment (e.g. fixing it on a pole) and robustness against atmospheric agents (Fig. 3).

The optical image provided by the video camera is overlapped onto the acoustic image, displayed as a color coded image, encoding at each pixel the intensity of the sound coming from a given direction, according to Section 2.1. An example of acoustic- optic image is shown in Fig. 1.

The device is characterized by a working band of 200 Hz - 10 kHz with an acoustic frame rate of 12 frames per second. The maximum field of view is 90° in elevation and 360° in azimuth (tunable according to the video camera field of view). The optimized microphone layout and filtering procedure [6] provides an acoustic image resolution, measured at -3 dB, of 5° at 6400 Hz[3]. The relative dynamic range of the acoustic image (i.e. ability to visualize in the same frame two sources with a different intensity) is equal to 30 dB. Since the device is a passive sensor the range is not defined a priori because it depends on the intensity of the source. To give a concrete example, we were able to detect noise of highway traffic up to 500 meters away from the planar array. Notice that, if the acoustic sources are particularly far from the device and an accurate synchronization of audio and video is necessary, it is advisable to consider and compensate for the propagation time of the acoustic wave. Depending on the application, it is necessary to assess whether the overall delay in the resulting acoustic/optic image display is acceptable.

The computation of the acoustic map and the overlapping with the optical one is performed on board, in real time, by means of a System on Chip, equipped with a Field Programmable Gate Array (FPGA) processor. Therefore, the device can be used stand alone, simply connecting it to a display, or it can be connected to a security network by a standard Ethernet connection and remotely controlled.

Another characteristic of the device is the compliance with different kinds of cameras, including thermal and infra-red ones, that can easily substitute the optical one according to the required application.

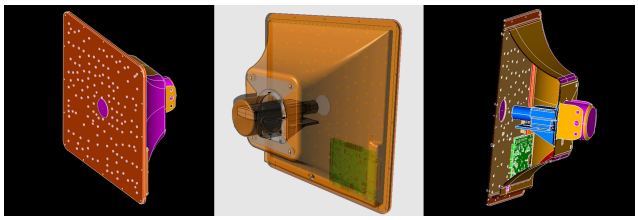

Figure 3. CAD model of the developed microphone array prototype.

3. Audio-video tracking

We briefly introduce the two acoustic tracking methods used in our experiments in Section 3.1 and 3.2 and we will compare them with a recent video tracking method presented in Section 3.3.

3.1. Kalman filter

The first acoustic tracking method considered is the Kalman filter [5]. The target state is a 4-dimensional vector: $\boldsymbol{X} = [\tilde{x}, \tilde{y}, \dot{\tilde{x}}, \dot{\tilde{y}}]$, where $\tilde{x}$ and $\tilde{y}$ are the image coordinates of the target and $\dot{\tilde{x}}$ and $\dot{\tilde{y}}$ are its velocity components, again on the image. The measurement is a 2-dimensional vector $\boldsymbol{Z} = [\bar{x}, \bar{y}]$. The matrix F describing the dynamic model of the state and the measurement matrix H have been set as

[3]The main lobe of the beam pattern decreases at -3 dB with respect to its maximum at an elevation angle of 5°.

follows:

$$F = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{5}$$

where $dt = 1/5$. The dynamical model is quite simple, assuming constant target velocity, while the measure is set equal to the coordinates of the target in the image.

The noise covariance matrices on the state Q and on the measure R are diagonal and set as follows:

$$Q = \begin{bmatrix} \sigma_p^2 & 0 & 0 & 0 \\ 0 & \sigma_p^2 & 0 & 0 \\ 0 & 0 & \sigma_s^2 & 0 \\ 0 & 0 & 0 & \sigma_s^2 \end{bmatrix} \qquad R = \begin{bmatrix} \sigma_o^2 & 0 \\ 0 & \sigma_o^2 \end{bmatrix} \tag{6}$$

where σ_p^2, σ_s^2 and σ_o^2 are the position, velocity and measure variance respectively.

The measurement $\boldsymbol{Z}$ is given by the image coordinates of the maximum peak in the acoustic image.

3.2. Particle filter

The Kalman filter is optimal under the assumption of Gaussian additive noise in the measurements. However, when interfering sources are present or the environment is highly reverberant this assumption is grossly violated. In fact, the acoustic image may show more than one peak and often the highest one do not correspond with the location of the source of interest. Hence, taking the coordinates of the maximum of the acoustic image as measurement may lead to high inaccuracies in the tracking performance. For this reason we rely on the particle filter in which the posterior probability density of the state is approximated as a sum of weighted samples, named particles. The dynamic of each particle follows the same model adopted for the Kalman filter, while the weights are updated and particles resampled according to the likelihood function $p(\mathbf{Z}|\mathbf{X})$ extracted from the acoustic map. More details on the particle filter implementation can be found in [8]. In a reverberant environment, characterized by a continuous target, such as voice, moving around and impulsive interfering sources (e.g. footsteps) only the peak location due to the target is temporally consistent across different frames, while reverberation and interferers are not [17]. Thus, particles will tend to cluster around the true peak, while other regions will be poorly sampled. If in a single frame an undesired peak arises, it will influence only a few particles and the state position estimation will not substantially change.

Concerning the likelihood function, we exploited the full normalized acoustic map, instead of taking just its maximum, as follows:

$$p(\mathbf{Z}|\mathbf{X}) = exp\left(\frac{M_a(\tilde{m},\tilde{n})}{\sigma_N^2}\right) \tag{7}$$

where $\tilde{m}$ and $\tilde{n}$ are the pixel indexes whose corresponding image coordinates are the closest the first two element of the state $\mathbf{X}$ (i.e. $(\tilde{x},\tilde{y})$), and σ_N^2 is the expected noise variance. In practice each particle weight is updated with the exponential of the acoustic map evaluated at the particle coordinates. The exponentiation enhances the likelihood sharpness around its peaks dampens the ground values around the value 1. Therefore regions far from every source will assume very uniform low values, while the peak corresponding will be increased. As a consequence, when some source is active the significant particle weights will be concentrated around the true location in the image, whereas if no source is active the weights will be close to 1 over all the image. The noise variance σ_N^2 tunes the sharpness of the likelihood function.

3.3. TLD video tracking

Several visual tracking methods combine tracking, learning and detection in a single framework. In [18], an offline trained detector is used to validate the trajectory output by a tracker and if the trajectory is not validated, an exhaustive image search is performed to find the target. Other approaches integrate the detector with a particle filtering [10] framework. In the recent years, adaptive discriminative trackers [3], [9], [21] also have the capability to track, learn and detect. These methods perform tracking by an online learned detector that discriminates the target from the background. In other words, a single process represents both tracking and detection.

Unlike them a very effective algorithm is TLD (a.k.a. Predator) where tracking and detection are independent processes that exchange information using learning [11]. We tested this visual tracking method in our experiments to compare it to the previous introduced audio tracking approaches. We gave as input to the tracker the sequence of gray-scale frames acquired by the optical camera embedded in the proposed device.

4. Experiments

We acquired three audio - video sequences[4]. The first, and second one, about 2 minutes long, are taken in a moderately reverberant environment i.e. a room with reflective floor, ceiling and furniture and walls partially covered with anechoic panels. The third one is taken outdoor from a terrace, looking at a road about 50 m far from the device. The three sequences present increasingly challenging conditions for the audio tracker. In the first one, the goal is the tracking of a drone flying in the room, whose propellers are the only active audio sources present in the scene (see Fig. 4). In the second one the goal is the tracking of the face of a

[4]The dataset is publicy available at: http://www.iit.it/en/pavis/datasets/DualCam.html

speaker moving in the room (see Fig. 5). In this case, also disturbing audio sources, generated by other people moving in the room (e.g. footsteps, body movements) are present in the scene. Finally in the outdoor sequence the goal was the tracking of a motorbike. In this case the target is quite far from the device and many disturbing sources are present, including wind flurries causing trees movements, an highway at about 500 m from the camera in the upper zone of the image, air conditioning units and people speaking on the terrace (see Fig. 6). Since we are interested at single target tracking we concatenated a set of 9 sub-sequences (each one about 10 seconds long) where just one motorbike rode the monitored street.

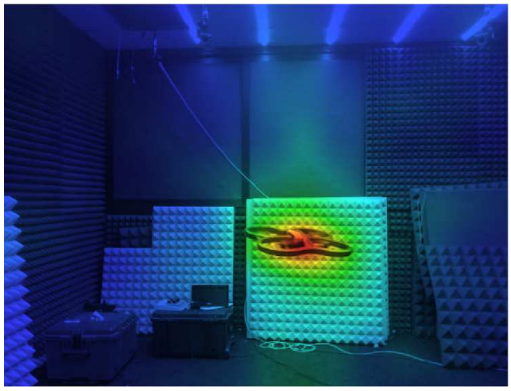

Figure 4. An example audio/video frame extracted from the drone-sequence.

Figure 5. An example of audio/video frame extracted from the voice-sequence. In the acoustic map are visible the target sound (the voice) and some other sounds generated by the people movement.

For the outdoor sequence, the beamforming is not focused because the source was in far-field, while for the two indoor sequences the focalization is set to 2.5 meters since the sources were in near-field for a considerable range of the frequencies. Acoustic images taken as input to the audio tracker were generated on the frequency range $500 - 6400$ Hz. Acoustic images were evaluated at 48×36 pixels and

Figure 6. An example of audio/video frame extracted from the outdoor-sequence. In the upper-right part the zoomed image of the target motorbike is displayed. Disturbing sounds are clearly visible in the upper left zone corresponding to the highway.

subsequently resized to 640×480 pixels. The acoustic tracking algorithms are initialized at the image center while the video tracking one, requiring the initial target appearance, is initialized at the target ground truth position in the first frame of the sequence. This difference accounts for the fact that, for audio tracking, ground truth initialization by a human operator may be difficult to obtain, especially in presence of multiple interfering sources. Tracking performance was calculated as the average Euclidean distance over frames between ground truth and estimated target position on the image plane [5].

4.1. Drone-sequence

The average error for the drone sequence is reported in Fig. 7 for the Kalman filter, function of position and measurement variances σ_p^2 and σ_o^2 (see Eq. 6), and in Fig. 8 for the particle filter, function of noise variance σ_N^2 (see Eq. 7) and number of particles. The lowest error for Kalman filter, i.e. 17.5 pixels, is slightly better than the corresponding one for particle filter (21 pixels). Moreover Kalman filter seems to be more robust toward the tuning of its parameters σ_p^2 and σ_o^2, showing only a slight increase of the error till 19.5 pixels. On the contrary the particle filter appears to be more sensitive to the number of particles employed and the range of values of σ_N^2. The overall better performance of Kalman filter is explanable considering the simplicity of the task from an acoustic point of view: just one audio source, with a very stable energy given by the drone propellers, is present in the scene, hence the assumptions under which Kalman filter is optimal are easily fulfilled. Notice that the value of 17.5 pixel is only slightly above the expected error in manual ground truth annotation. The performance changes dramatically when relying on the TLD tracker ap-

[5] The TLD tracker may yield no output in frames where the target is not detected: in these cases we set the estimated position of the target equal to the last valid output.

plied on the optical image. Due to the abrupt changes in appearance of the drone, TLD loses the target just after one or two frames and is not able to recover. This result does not change tuning the TLD parameters, such as the size of the target bounding box, making quantitative results meaningless for this sequence.

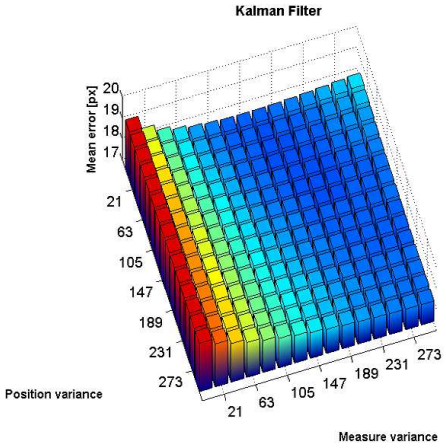

Figure 7. Average error in pixels versus σ_p^2 and σ_o^2 for the drone sequence and Kalman filter.

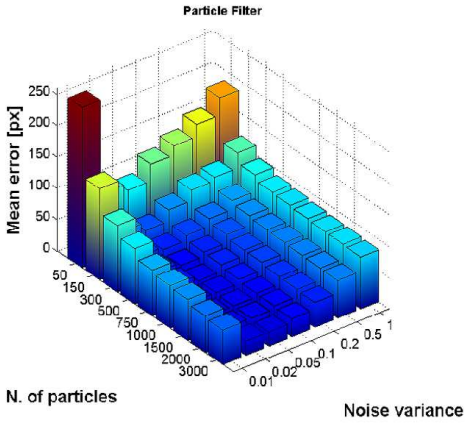

Figure 8. Average error in pixels versus σ_N^2 and number of particles for the drone sequence and particle filter.

4.2. Voice-sequence

As mentioned before, the speaker tracking is a more challenging task with respect to the drone one from an audio perspective, due to both the intrinsic variability in the voice energy and the presence of impulsive noises such as the footsteps of the people in the room. This explains the comparatively worse results for audio tracking with both Kalman and particle filter, reported in Fig. 9 and Fig. 10, respectively. Differently from the drone sequence, the best result is achieved by the particle filter with a minimum error of 45 pixels, while the best result of Kalman almost doubles such error. The better result of particle filter is coherent with the non-gaussianity of measurement noise caused by spurious peaks, due to disturbances appearing in the acoustic image far away from the target position. While Kalman filter takes the maximum of the acoustic map as measurement, therefore shifting quite suddenly toward the wrong peak, particle filter may almost ignore peaks appearing in isolated frames. In fact the whole acoustic map is taken as input and non- uniformly sampled by the particles that are clustered around the true target position; consequently in the next prediction step just a few particles will be influenced by the spurious peaks with a negligible effect on the overall state estimation.

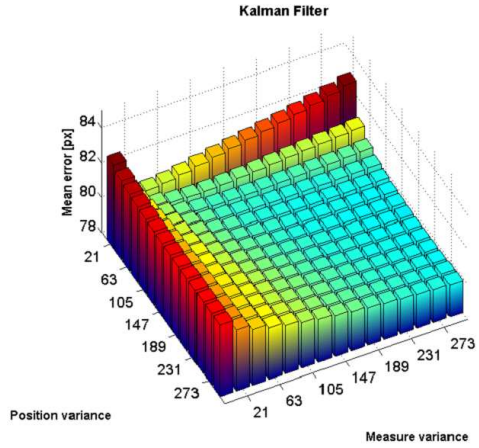

Figure 9. Average error in pixels versus σ_p^2 and σ_o^2 for the voice sequence and Kalman filter.

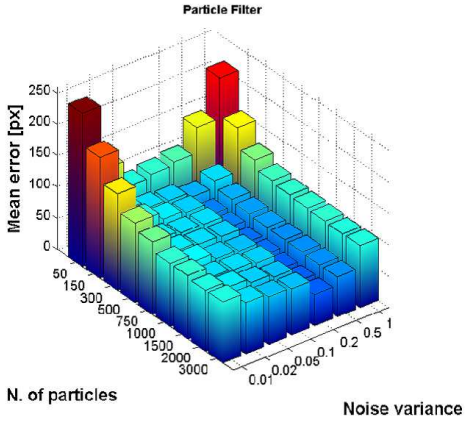

Figure 10. Average error in pixels versus σ_N^2 and number of particles for the voice sequence and particle filter.

Concerning the TLD tracker on the optical image, average error versus the bounding box size of the target is reported in Fig. 11. The lowest error is of about 63 pixels, for a bounding box size of 45×45 pixels and the error is kept

below 80 pixels in the bounding box range 15×15 - 55×55. This reasonable result, better than the one obtained with Kalman filter on the audio image, is due to the slowly varying appearance of the speaker face and the limited amount of occlusion. However the result is significantly overcome by audio tracking based on particle filtering, demonstrating also in this case the usefulness of audio imaging, even in presence of disturbing audio sources.

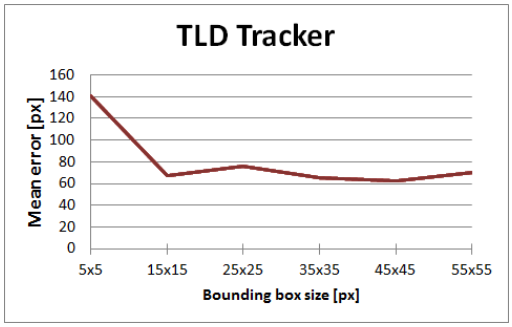

Figure 11. Average error in pixels versus the size of the target bounding box for the TLD tracker on the optical image.

4.3. Outdoor-sequence

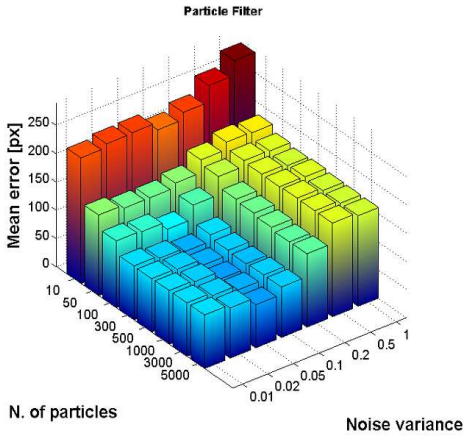

Figure 12. Average error in pixels versus σ_N^2 and number of particles for the motorbike sequence and particle filter.

The average error for the motorbike sequence, obtained with the particle filter, is displayed in Figure 12. The best performance (76 pixel of error) is obtained with the highest number of particles and a noise variance $\sigma_N^2 = 0.05$. Even if the result is worse than in the previous setups, mainly due to the large amount of disturbing audio sources present in the scene, still the tracking performance is qualitatively acceptable, with the estimated target position being in the neighbourhood of the ground truth in the great part of frames and just occasional drifting toward the highway noise. Results for the Kalman filter are in this case clearly worse, achieving in the best case an error of 95 pixels. This fact is easily explainable considering that the highest peak does not correspond to the target position in most of the frames.

Similarly to the drone sequence, the TLD tracker on the optical image is not able to follow the target at all, losing it after just the first frame. The cause of this behaviour lies both in the low resolution of the motorbike target image, due to its considerable distance from the camera, and the huge amount of occlusion, due to trees and walls completely hiding the motorbike in many frames. Differently, audio imaging is little affected by occlusions, due to the nature of sound waves propagation, thus explaining the dramatically superior results achieved in this sequence. The comparative results on the three sequences are resumed in Table 1. In two sequences the visual tracker completely fails, despite using a very recent and performing algorithm. Moreover, in all the sequences acoustic tracking outperforms the visual one, either employing Kalman or particle filters.

	Kalman filter	Particle filter	TLD tracker
Drone	**17.5**	21	*
Voice	80	**45**	63
Outdoor	95	**76**	*

Table 1. Lowest average tracking error in pixels for the three setups and the three tracking algorithms.

5. Conclusions and future works

We proposed a new, low cost, acoustic-optic imaging device specifically taylored for automated surveillance. We demonstrated the device capabilities for tracking purposes on three different setups. Results show that audio imaging can solve tracking problems that cannot be handled by visual tracking, even with state-of-art algorithms. The current research will be developed along different directions. First of all, audio and video imaging can be fused together in a hybrid tracking algorithm, taking advantage of the complementary information brought by the two modalities. Moreover, in order to discriminate between target of interest and disturbing sources the spectral signature of audio signals associated to each pixel of the acoustic image can be extracted and fed to machine learning algorithms. Finally, we will investigate the possibility to couple our device with a Pan, Tilt and Zoom (PTZ) camera in a master-slave modality: once an anomalous audio event is detected and localized, the PTZ camera can be automatically steered and zoomed toward the area of interest acquiring high resolution visual information. Having such high-resolution data is of extreme importance in order to make Computer Vision algorithms feasible for real-world applications.

References

[1] http://www.gfaitech.com/de/products/akustische-kamera.html. 2

[2] http://www.norsonic.com/en/products/acousticcamera. 2

[3] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online Multiple Instance Learning. In *Computer Vision and Pattern Recognition*, pages 983–990. IEEE, June 2009. 2, 5

[4] V. Cevher, R. Velmurugan, and J. H. McClellan. Acoustic multitarget tracking using direction-of-arrival batches. *Signal Processing, IEEE Transactions on*, 55(6):2810–2825, 2007. 3

[5] S. Y. Chen. Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics*, 59(11):4409–4420, 2012. 4

[6] M. Crocco and A. Trucco. Design of superdirective planar arrays with sparse aperiodic layouts for processing broadband signals via 3-d beamforming. *Audio, Speech, and Lang. Proc., IEEE/ACM Trans. on*, 22(4):800–815, 2014. 3, 4

[7] M. F. Fallon and S. Godsill. Acoustic source localization and tracking using track before detect. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1228–1242, 2010. 2

[8] N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993. 5

[9] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 234–247, Berlin, Heidelberg, 2008. Springer-Verlag. 2, 5

[10] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998. 5

[11] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1409–1422, July 2012. 2, 5

[12] E. A. Lehmann and A. M. Johansson. Particle filter with integrated voice activity detection for acoustic source tracking. *EURASIP Journal on Applied Signal Processing*, 2007(1):28–28, 2007. 2

[13] F. Pernici and A. Del Bimbo. Object tracking by oversampling local features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2538–2551, Dec 2014. 2

[14] A. Plinge and G. Fink. Multi-speaker tracking using multiple distributed microphone arrays. In *Acoust., Speech and Sig. Proc. (ICASSP), 2014 IEEE Int. Conf. on*, pages 614–618, May 2014. 2

[15] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: an experimental survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1442–1468, 2014. 2

[16] H. Van Trees. *Detection, Estimation, and Modulation Theory, Optimum Array Processing*. Detection, Estimation, and Modulation Theory. Wiley, 2002. 3

[17] D. B. Ward, E. Lehmann, and R. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *Speech and Audio Proc., IEEE Trans. on*, 11(6):826–836, 2003. 2, 5

[18] O. Williams, A. Blake, and R. Cipolla. Sparse bayesian learning for efficient visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1292–1304, Aug 2005. 5

[19] K. Wu, S. T. Goh, and A. W. Khong. Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 365–369. IEEE, 2013. 2

[20] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[21] Q. Yu, T. B. Dinh, and G. G. Medioni. Online tracking and reacquisition using co-trained generative and discriminative trackers. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *ECCV (2)*, volume 5303 of *Lecture Notes in Computer Science*, pages 678–691. Springer, 2008. 2, 5