GENERATING TRAIN SIDE VIEWS FROM VIDEO SEQUENCES FOR MICROPHONE ARRAY PASS-BY MEASUREMENTS

Johannes Stier and Michael Beitelschmidt
{Johannes.Stier, Michael.Beitelschmidt}@tu-dresden.de
Technische Universität Dresden, Institut für Festkörpermechanik,
Professur für Dynamik und Mechanismentechnik
Marschnerstraße 30, 01307 Dresden

ABSTRACT

A significant field of application for microphone arrays is the pass-by measurement of moving objects, mainly trains. In a former article, a special time-domain beamforming algorithm has been introduced which is able to evaluate whole pass-by measurements with respect to computational performance and memory. The evaluation results in a sound pressure level mapping of the train passing. But this mapping only shows the sound sources and their positions represented by coordinate values resulting in interpretation difficulties.

A possible solution is the overlay of the sound pressure level mapping with a sketch or a picture of the whole train side view. Since this kind of picture is only available in very few cases or only for single parts of the train, an approach was developed which allows the picture generation of a train side view independently from its composition. In addition to the array measurements, the train's passing is recorded with a video camera simultaneously. By analyzing the video frames with the SIFT algorithm and combining them based on the gained information, a side view picture of the whole train can be created.

The approach presented, including its advantages and requirements, is described in detail. Additionally, further developments are exposed and the capabilities of this approach are demonstrated with an example taken from a variety of measurements completed so far.

1 INTRODUCTION

In 2010, a special time-domain beamforming algorithm was introduced [5]. Being capable to evaluate whole pass-by microphone array measurements, the algorithm creates a sound pressure level mapping (SPLM) of a passing train. This mapping can be used to identify dominating sound sources with their spatial position in two dimensional space. Depending on their value, the sound pressure levels calculated get a color assigned. Figure 5.(b) illustrates an example for

a SPLM. Although the spatial position of the sound sources revealed is a result of the beamforming evaluation, the train parts being responsible for the sound can only be guessed. The exact identification of the sound sources could be simplified, if an image of the whole train's side view would be available.

One way to create the image required is taking a picture of the passing train. But in many cases, this is nearly impossible or very difficult, because of the enormous length of the train. In general, when using a regular camera, only partial sections of the train can be displayed, e.g. single cars or the locomotive. Additionally, this requires a wide-angle lens and a very small focal length has to be used, which leads to distortions in the images taken. Thus an overlay with the whole SPLM seems to be impossible, which decreases the benefit of the developed beamforming algorithm.

For documentation purposes, there is usually a video camera available for recording a train's passing simultaneously to the microphone array measurement. The resulting videos contain the whole image data necessary for an overlay with the SPLM. Only an approach for combining the picture data to an image of the train's entire side view is missing. For that reason, an algorithm was developed making this image available. Applying this algorithm to a video of a train passing results in a complete side view image (SVI).

The algorithm presented is implemented in Matlab, because of its wide toolbox variety. A special image processing algorithm is used to create the base for combining the video's picture data. To reduce the programming effort, an available implementation of this algorithm was used.

2 THE BASIC IDEA

The input of the algorithm to be developed is the video of the side view of a passing train. A video consists of a certain set of single images, called frames, which are recorded with a defined frame rate. The most common frame rate is 25 frames per second (fps). Assuming a recorded video to be 20 seconds of length, this results in 500 frames, each containing partial sections of the train's side view. The task to be solved is the determination of the relation between adjacent frames and their linking to an entire SVI, which represents the algorithm's basic idea.

Creating panorama images on the base of single pictures, e.g. taken from a scenic view of a mountain, is quite similar to the main idea described. The panorama creation requires data depicting the stitching of the single images, which is often determined by applying an algorithm called SIFT (Scale Invariant Feature Transform) [1, 2]. Other fields of application of SIFT are the visual detection of objects in images [3] or the localization and mapping in robotics [4].

The SIFT algorithm analyzes an image and detects distinctive features, so called keypoints. They are invariant to scale and rotation, and partially invariant to illumination and to the 3D camera viewpoint [3]. Using SIFT to analyze two images which are partially overlapping, thus containing sections of the same scene, will produce a set of keypoints each. Because these keypoints are invariant to the factors mentioned above, there will be a set of common keypoints in the area of overlapping. Matching the common keypoints results in keypoint pairs, which exactly describe the geometrical relation between the images. Based on this description the images can be stitched to a combination of both images.

If the video frames are considered to be the single images of a panorama to be created, the approach presented will be applicable as well. In this case, two adjacent frames are analyzed by

SIFT to identify the keypoints. After the keypoints have been matched, the frames' relation can be described and it is possible to stitch them. Applying these steps to all video frames, a whole panorama image of the train's side view can be created.

Generally, the description of the geometrical relation between two images requires the consideration of translation, rotation, scaling and shearing. In this case, the spatial transformation matrix describing this relation has to be approximated by an estimation based on the keypoint pairs. This is quite complex and challenging due to the determination and correction of distortions.

To keep the algorithm as simple as possible, some assumptions and restrictions are defined. Reducing the geometrical relation between the video frames to one dimension leads to a first simplification. This means, the train travels in horizontal direction through the image section recorded by the camera. Thus, only the displacement (translation in x-direction) of adjacent frames needs to be determined, and the determination of the transformation matrix is unnecessary. Furthermore, the following assumptions and simplifications have to be regarded:

1. The video camera is perfectly aligned and there are no distortions due to focal length. This is achieved by the image section's horizontal edges being parallel to the rail and the overhead lines. (The train recorded can be considered as a rectangle moving through a rectangular image section, with its edges being parallel to the image section edges.)

2. The train is traveling with constant velocity. Actually, because of the beamforming algorithm's requirement on constant velocity, this is not an explicit constraint.

3 FROM MOVIE TO SIDEVIEW – THE ALGORITHM

3.1 Determining the Displacement of Adjacent Frames

The video being the origin of the side view picture has a resolution of $W \times H$ Pixel (Px), with the width W (number of columns) and the height H (number of rows). It consists of N frames F_n ($n = 1 \ldots N$) with every video frame being a raster graphics and having the same resolution as the video. If no other unit is indicated, the unit used in the following sections is Px (Pixel).

As said previously, the first step of the SVI creation is the analysis of two adjacent frames F_n and F_{n+1} with the SIFT algorithm. This results in a set of keypoint pairs containing the identified and matched keypoints. Thus the geometrical relation of the frames is completely determined. In the set of keypoint pairs, there are three different types: pairs belonging to the train, pairs belonging to the background and even non-matching pairs. For the calculation of the horizontal displacement of the frames, the keypoint pairs belonging to the train have to be selected certainly.

There are several selection methods contained in the external implementation of the SIFT algorithm. But their application is not possible. These methods would select the keypoint pairs of the background, because they are very stable since there are no changes in background during recording. For that reason, new selection methods have to be developed. The result is a quite simple three-step selection process, which will be introduced below.

The starting point is a set of J keypoint pairs of the two adjacent frames, consisting of the keypoints $K_n^j = \left(x_n^j, y_n^j\right)$ of frame F_n and $K_{n+1}^j = \left(x_{n+1}^j, y_{n+1}^j\right)$ of frame F_{n+1} with $j = 1 \ldots J$.

Initially, the distance in x-direction $d^j_{x,n}$ and y-direction $d^j_{y,n}$ is calculated based on the following equations:

$$d^j_{x,n} = \left| x^j_{n+1} - x^j_n \right|, \tag{1}$$

$$d^j_{y,n} = \left| y^j_{n+1} - y^j_n \right|. \tag{2}$$

Afterwards, the three-step selection method is applied to choose the required set of keypoint pairs:

Step 1: *y-threshold selection*

The distance $d^j_{y,n}$ of the J keypoint pairs is compared to a given threshold ε_y:

$$d^j_{y,n} < \varepsilon_y. \tag{3}$$

K keypoint pairs are selected, that satisfy the condition above.

Assumption: The train is moving perfectly in horizontal direction through the image section. For all keypoint pairs of the train the distance $d^j_{y,n}$ is approximately zero ($d^j_{y,n} \approx 0$).

Step 2: *x-threshold selection*

The distance $d^k_{x,n}$ of the K keypoint pairs is compared to a given threshold ε_x:

$$d^k_{x,n} > \varepsilon_x. \tag{4}$$

L keypoint pairs are selected satisfying the condition above.

Assumption: The train is moving with the frame background being constant. For all keypoint pairs of the background follows $d^k_{x,n} \approx 0$.

At this point, all keypoint pairs belonging to the train are selected. But still keypoint pairs with non-matching keypoints can exist. Thus another selection step is necessary.

Step 3: *Statistical selection*

M keypoint pairs are selected according to the statistical distribution of $d^l_{x,n}$ out of the L keypoint pairs left. Those keypoint pairs are selected, whose x-distance lies within the range $\pm\delta_x$ around the average value $\overline{d_{x,n}}$:

$$\overline{d_{x,n}} - \delta_x \leq d^l_{x,n} \leq \overline{d_{x,n}} + \delta_x, \tag{5}$$

Assumption: The train is moving with constant velocity through the image section. As a result, the x-distance of the matching keypoints has to lie in a certain range, when neglecting distortions and the aligning accuracy of the camera.

The result of the entire selection process is a set of M keypoint pairs belonging to the train. All other pairs, those belonging to the background and those with non-matching keypoints, are eliminated.

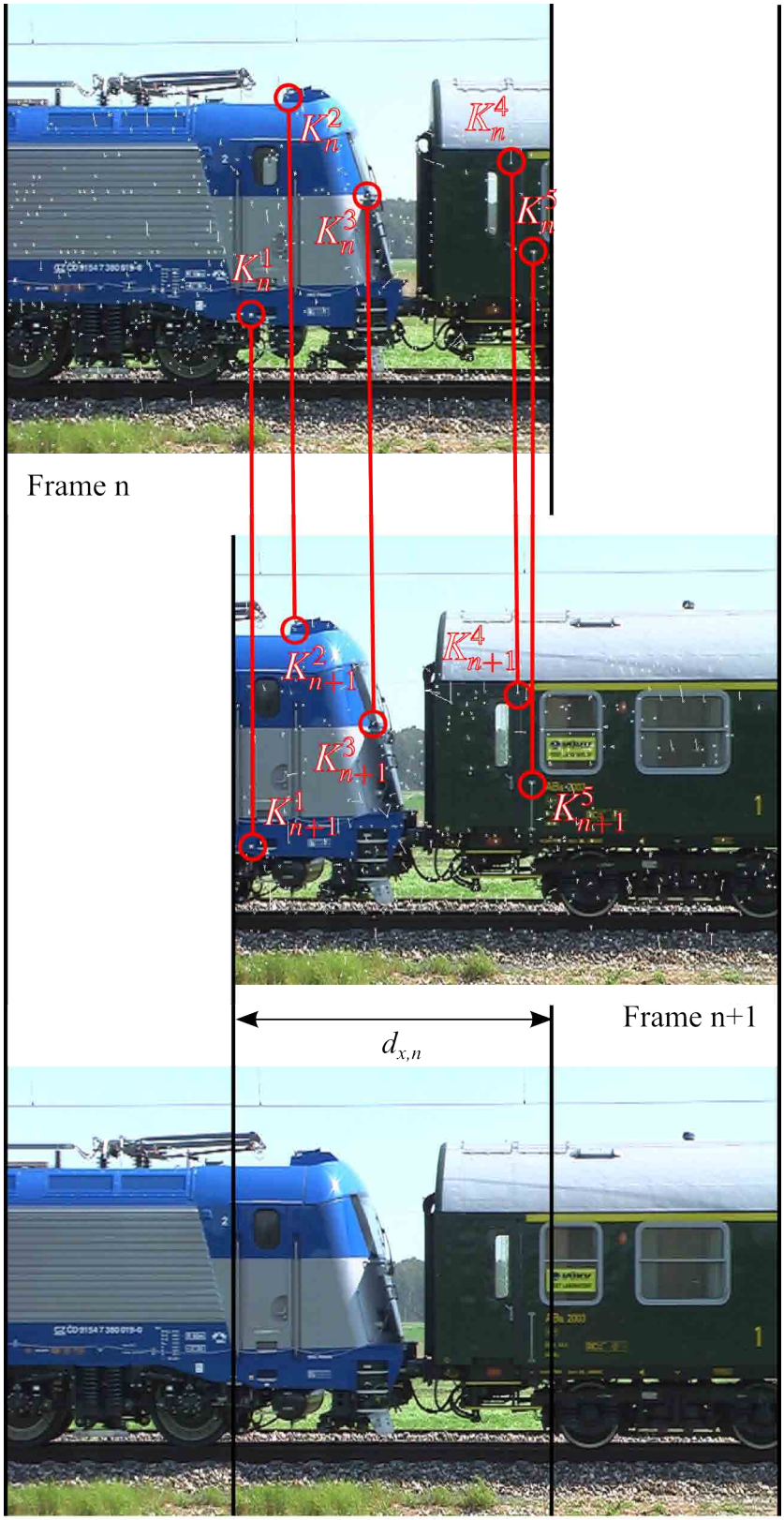


Figure 1: Stitching two adjacent frames F_n and F_{n+1} according to their distance $d_{x,n}$

Finally, the distance $d_{x,n}$ of the adjacent frames F_n and F_{n+1} is calculated by averaging all M distances $d_{x,n}^m$ of the keypoint pairs selected:

$$d_{x,n} = \overline{d_{x,n}}. \tag{6}$$

Consequently, the displacement of the adjacent frames is now completely determined according to the simplifications made concerning the geometrical relation. In the next step, all frames F_n can be stitched based on the distances $d_{x,n}$ calculated to create the SIV.

Figure 1. shows an example which illustrate the displacement of two adjacent frames depending on their distance $d_{x,n}$. The white points indicate the keypoints detected by the SIFT-algorithm. In every frame, five matching keypoints (K_n^1/K_{n+1}^1 to K_n^5/K_{n+1}^5) are marked.

3.2 Stitching Adjacent Frames

After determining the displacement of adjacent frames, these frames can be stitched to get a combined image. The approach applied will be described using the example of the adjacent frames F_n and F_{n+1}, because it is much easier to follow.

Initially, an empty image I_{res} of height $H_{res} = H$ and width $W_{res} = 2 \cdot W - d_{x,n} + 1$ is created, yielding from the width W of the frames and the their distance $d_{x,n}$. From column 1 to column $W - d_{x,n}$ the data of frame F_n from column 1 to column $W - d_{x,n}$ is added. Afterwards, the remaining columns that are still empty are filled with the entire data of frame F_{n+1} from column $W - d_{x,n} + 1$ to column W_{res}. The result is the stitched image of frame F_n and F_{n+1}.

An example of this procedure is illustrated in Fig. 5. In this case, the two frames F_n and F_{n+1} are stitched according to their distance $d_{x,n}$.

3.3 Creating the Side View

In the previous section, the stitching of two frames was described to demonstrate the basic stitching approach for reasons of simplification. The creation process of the entire SVI is very similar.

Before creating the SVI, all $N-1$ frame pairs of the N video frames are analyzed with the SIFT algorithm. Subsequently, the approach to determine the distances $d_{x,n}$ of adjacent frames is applied according to subsection 3.1, resulting in $N-1$ distance values.

For quality reasons, another validation step was introduced to increase the reliability of the SVI creation process. To decrease the influence of non-matching keypoint pairs on the distances $d_{x,n}$, and thus on the SVI, the $d_{x,n}$ are partially corrected. Assuming a constant train velocity, all $d_{x,n}$ must lie within a certain range. The center point of this range is denoted by the median value $\widetilde{d_{x,n}}$ of all $d_{x,n}$ and the range is defined by $\pm\delta_d$:

$$\widetilde{d_{x,n}} - \delta_d \leq d_{x,n} \leq \widetilde{d_{x,n}} + \delta_d. \tag{7}$$

All R displacement values lying within this range are declared with $d_{x,in,r}$ ($r = 1 \ldots R$), and all other $S = N - R - 1$ values with $d_{x,out,s}$ ($s = 1 \ldots S$). Afterwards, the values of the distances $d_{x,out,s}$ are reset to the average value of the $d_{x,in,r}$:

$$d_{x,out,s} = \overline{d_{x,in,r}}. \tag{8}$$

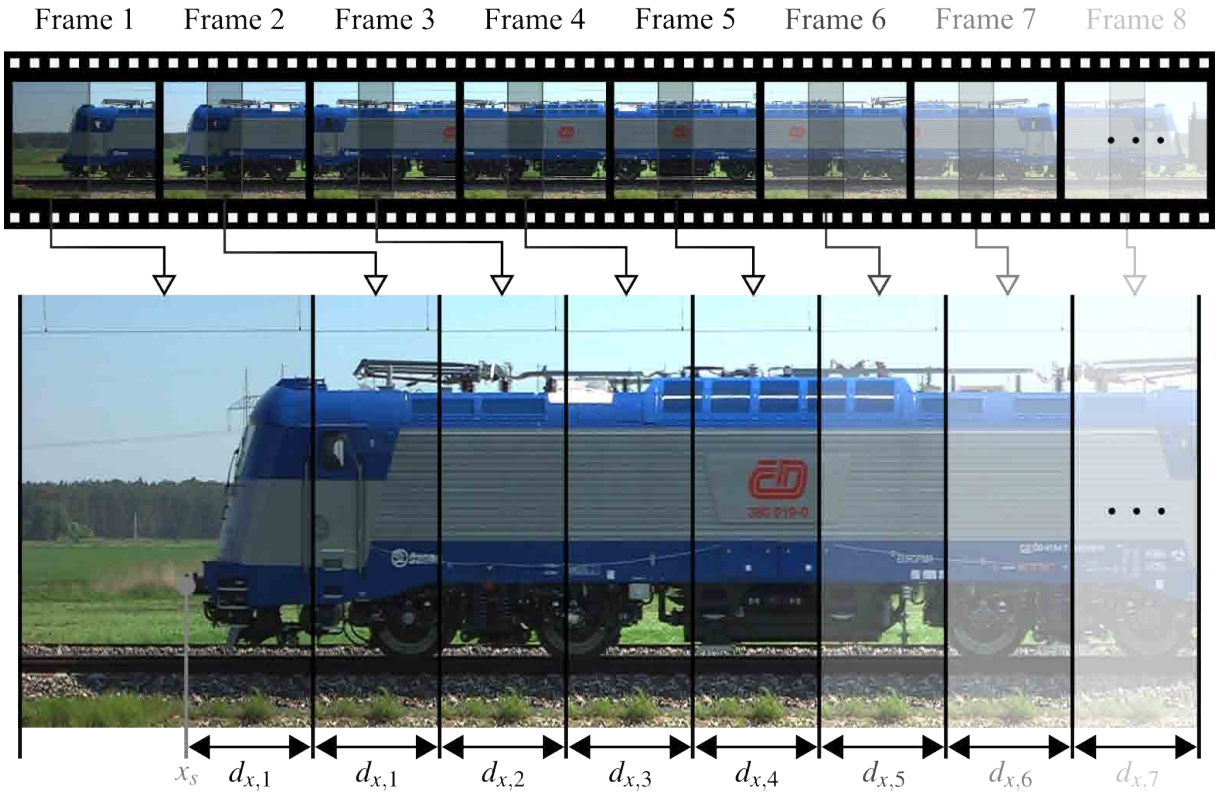


Figure 2: Creating a side view image by appending frame stripes to the first frame

By this mean, all requirements for the SVI creation are satisfied so far. The creation process starts with the manual selection of a horizontal position x_s in the first frame F_1, which denotes the very first part of the train. In general, this is the buffer of the train. By choosing this mark, the area of the frames being used for the SVI creation can be chosen. For least distortions, this mark should be conveniently in the middle of the frames.

Subsequently, an empty image $I_{res,1}$ of width $W_{res,1} = x_s + d_{x,1}$ is created. The height $H_{res} = H$ is determined by the height of the frames, and will not change during the creation process. Now, the content of frame 1 from column 1 to column $x_s + d_{x,1}$ is copied into $I_{res,1}$ to the columns with the same index. Expecting all $d_{x,n}$ being within the same range, and they are referring to the validation step at the beginning, the approach implemented can be considered as stitching frame 1 to a virtual frame 0. By considering the distance value $d_{x,1}$ for their displacement, a stripe of width $d_{x,1}$ of frame 1 is used. Actually, according to the procedure described in subsection 3.2, the columns in the range $[x_s, x_s + d_{x,1}]$ of $I_{res,1}$ would be replaced by a stripe of frame 2 with the column range $[x_s - d_{x,1}, x_s]$. But depending on the selected position of x_s, distortions in the resulting image $I_{res,1}$ would become much higher. Additionally, the image data of frame 1 would be useless.

The image $I_{res,1}$ denotes the starting point for the SIV creation steps remaining. All $i = 2 \ldots N$ frames left can be appended to $I_{res,1}$ according to the following approach:
A stripe of width $d_{x,i-1}$ starting at column x_s is extracted from frame i. This stripe is then just

appended to the resulting image $I_{res,i-1}$. The result is the image $I_{res,i}$ of width

$$W_{res,i} = x_s + \sum_{l=2}^{i} d_{x,l-1}. \tag{9}$$

Repeating this procedure for all frames produces an image $I_{res,N}$, which includes the train side view.

Figure 2. illustrates the approach described above. In this case, the creation process is demonstrated for the first eight frames. Starting with the special treatment of frame 1, stripes of width $d_{x,n}$ are cut out from the $n = 2 \ldots 8$ frames following, and appended to the previous created image part.

3.4 The Algorithm

In the previous three sections, the basic approach for creating a SVI was described. These steps are now be summarized by an algorithm description:

Frame Extraction

1. Extract all N frames from the video

Frame Analysis

- ⟳ Repeat for all $n = 1 \ldots N-1$ frame pairs
  2. Analyze frame pair n with the SIFT algorithm to get the J keypoint pairs
  - ⟳ Repeat for all $j = 1 \ldots J$ keypoint pairs
    3. Calculate the horizontal displacement $d^j_{x,n}$ and vertical displacement $d^j_{y,n}$ for the keypoint of pair j (*Eqn. (1), (2)*)
    4. Select L keypoint pairs by comparing $d^j_{x,n}$ and $d^j_{y,n}$ to the thresholds ε_x and ε_y (*Eqn. (3), (4)*)
  5. Select M keypoint pairs according to the distribution of the L $d^L_{x,n}$ left (*Eq. (5)*)
  6. Determine the average displacement $\overline{d_{x,n}}$ of the n-th frame pair (*Eq. (6)*)

SIV creation

7. Reset the displacement of the frame pairs whose displacement is outside of the given range to the average displacement (*Eq. (7)*)
8. Get the position of x_s in the first frame
9. Create the resulting image $I_{res,1}$ and copy the content of frame 1 into it
- ⟳ Repeat for all $i = 2 \ldots N$ frames
  10. Cut out a stripe of width $d_{x,i-1}$ from frame i starting at position x_s and append it to $I_{res,i-1}$ to create $I_{res,i}$
11. Output the final SIV $I_{res,N}$

4 APPLICATION

The algorithm described in section 3 was applied on a wide variety of videos from several microphone array measurements. Many different trains were recorded with a video camera and the algorithm always produced satisfying results. Although using a camera with a small resolution, the resulting resolution of the SVIs was always higher than the SPLM's resolution.

During the algorithm's development process, videos of trams were used for testing, since they do not differ from train videos and are easier to access. Figure 3. shows one of the first SVIs created.

A second example, taken from the repository of SVIs, is illustrated in Fig. 4. It is an SVI of a city train traveling around the city of Dresden. The train consist of four cars with one locomotive. It moved from right to left and traveled at a speed of 80 km/h. It is obvious that there are only a few distortions. This demonstrates the algorithms performance and the sufficient quality of the SVIs created.

The SVI of another regional train is shown in Fig. 5. In addition to the pure SVI, the image resulting from an overlay with a SPLM is displayed as well. This illustration exactly reveals the algorithm's purpose, because the bogie and the wheelsets can be identified as the dominating sound sources.

Despite the requirements established by the algorithm seem to be very restrictive, a suitable alignment of the camera can be achieved with some experience. Small existing distortions due to alignment problems can be reduced by down-sizing the SVIs, because an image smoothing arises from the interpolation method applied in this case. Moreover, recording a train traveling at low speed increases the quality of the SVI, because the displacement of the frames is small and motion blur caused by the train's movement can be reduced.

Figure 3: Full side view image of a tram in Dresden

5 CONCLUSION

An algorithm is presented that is able to create an entire side view image (SVI) from a recorded video of a passing train. The algorithm itself and further developed and implemented approaches were described in detail. Finally, the algorithm's application on several videos is demonstrated by selected examples. Although the algorithm restricts the video camera's positioning and aligning, it was illustrated that the algorithm can produce SVIs of satisfying quality.

Moreover, the algorithm only requires small computational effort and can be implemented very easily.

Nevertheless, the algorithm is applied successfully in many cases, the disadvantage caused by the restrictions on the camera's position and alignment cannot be neglected and implies some problems. Although taking the rail and the overhead line for the camera's alignment, there is sometimes not enough space to position the camera in the way necessary. Besides, if the camera is not positioned and aligned almost perfectly, or if the camera moves during recording, the SVI quality recedes quickly. Overlaying the SVI with the SPLM is done manually, because a method for synchronizing the video and microphone array measurement has not been developed so far. Thus the sound source assignment to train parts is an estimation.

For the reasons mentioned, an advancement of the algorithm is necessary and the problems caused by the restrictions justify a more complex and sophisticated algorithm. Instead of just determining the displacement of adjacent frames, approximating the transformation matrix describing the geometrical relation between frames and using it for stitching, will lead to a high increase in the algorithm's robustness and will raise the quality of the SVIs created. Moreover, developing an approach for the synchronization of the video recording and the microphone array measurement is a main goal for future work as well.

References

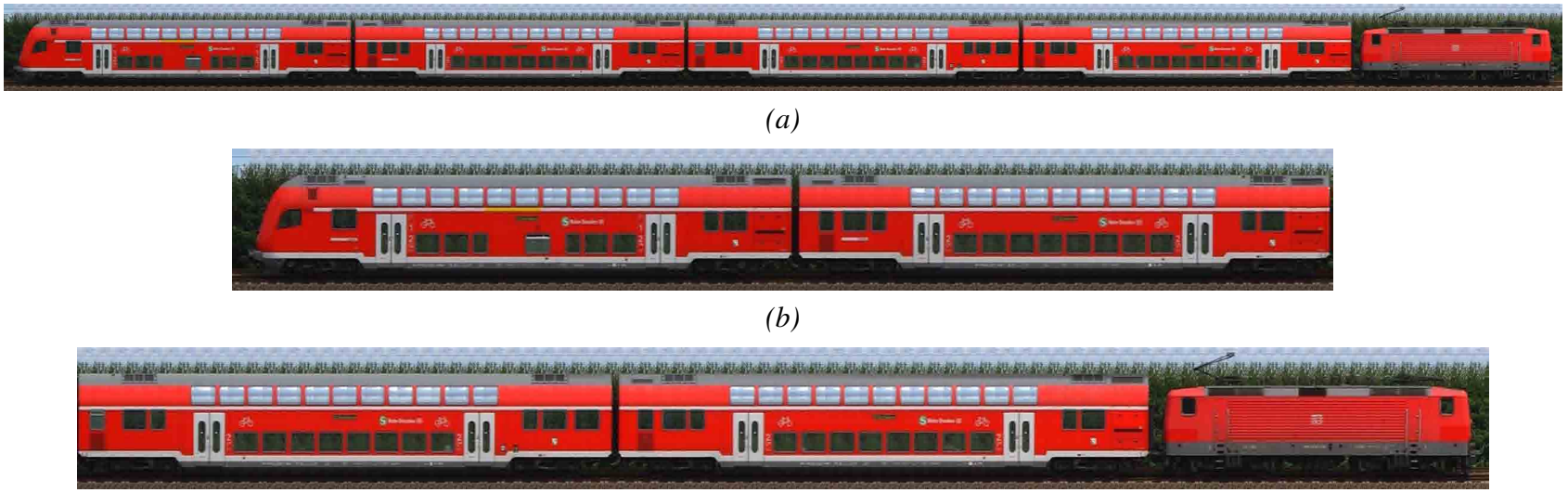
[1] M. Brown and D. G. Lowe. "Recognising panoramas." *International Conference on Computer Vision 2003*, 2, 1218–1225, 2003.

[2] M. Brown and D. G. Lowe. "Automatic panoramic image stitching using invariant features." *International Journal of Computer Vision*, 74, 16, 2007.

[3] D. G. Lowe. "Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, 60, 28, 2004.

[4] S. Se, D. G. Lowe, and J. Little. "Vision-based mobile robot localization and mapping using scale-invariant features." *IEEE International Conference on Robotics and Automation*, 2, 2051 – 2058, 2001.

[5] G. Zechel, A. Zeibig, and M. Beitelschmidt. "Time-domain beamforming on moving objects with known trajectories." In *3rd Berlin Beamforming Conference, Berlin*. 2010.

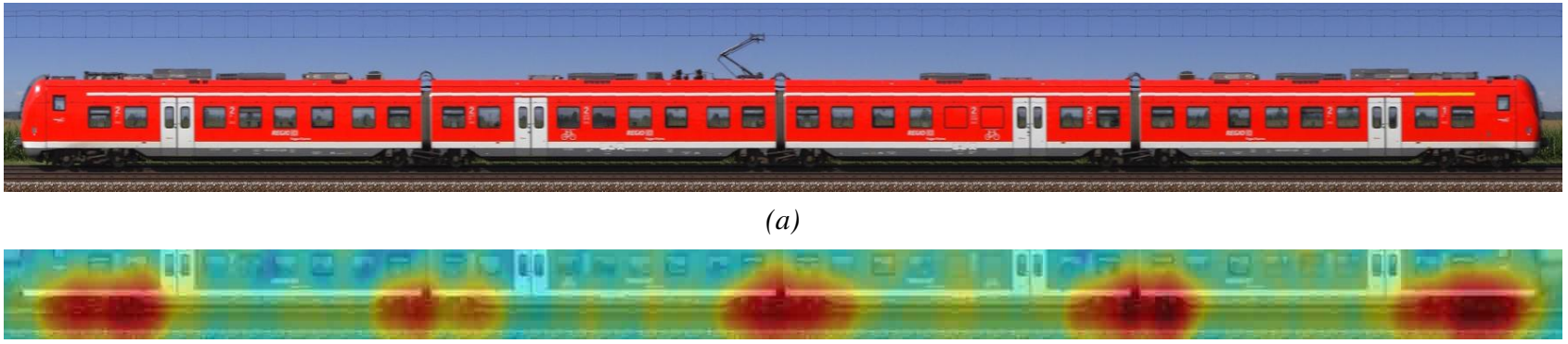


(a)

(b)

(c)

Figure 4: Full side view of a train (a), detailed illustration of the train's first two cars (b) and last two cars incl. the locomotive (c)



(a)

(b)

Figure 5: Full side view of a train (a), overlay of the side view with the sound pressure level mapping (b)