On detecting the playing/non-playing activity of musicians in symphonic music videos

Bazzica, A; Liem, CCS; Hanjalic, A

DOI
10.1016/j.cviu.2015.09.009

Publication date
2016

Document Version
Final published version

Published in
Computer Vision and Image Understanding

Computer Vision and Image Understanding 144 (2016) 188–204

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

On detecting the playing/non-playing activity of musicians in symphonic music videos

Alessio Bazzica*, Cynthia C.S. Liem, Alan Hanjalic

Delft University of Technology, Multimedia Computing Group, Mekelweg 4, 2628 CD Delft, The Netherlands

ARTICLE INFO

Article history:
Received 20 December 2014
Accepted 21 September 2015

Keywords:
Cross-modal analysis
Music information retrieval
Human-object interaction
Diarization
Clustering

ABSTRACT

Information on whether a musician in a large symphonic orchestra plays her instrument at a given time stamp or not is valuable for a wide variety of applications aiming at mimicking and enriching the classical music concert experience on modern multimedia platforms. In this work, we propose a novel method for generating playing/non-playing labels per musician over time by efficiently and effectively combining an automatic analysis of the video recording of a symphonic concert and human annotation. In this way, we address the inherent deficiencies of traditional audio-only approaches in the case of large ensembles, as well as those of standard human action recognition methods based on visual models. The potential of our approach is demonstrated on two representative concert videos (about 7 hours of content) using a synchronized symbolic music score as ground truth. In order to identify the open challenges and the limitations of the proposed method, we carry out a detailed investigation of how different modules of the system affect the overall performance.

1. Introduction

Rapidly developing multimedia technology has opened up new possibilities for bringing the full symphonic music concert experience out of the concert hall and into people's homes. New emerging platforms, like *RCO Editions*[1] and the *Berliner Philharmoniker's Digital Concert Hall*[2] are enriching audio-visual recordings of symphonic music performances to make them more informative and accessible offline, in a non-linear fashion and from multiple perspectives. Such platforms rely on the new generation of automatic music data analysis solutions. For instance, loudness and tempo can be estimated continuously over time and visualized as animations [8]. Notes can be detected and analyzed to reveal and visualize repeated parts of a piece [21]. Sheet music scores can be synchronized to the audio recording to allow users to follow the scores while listening to the music [2]. Furthermore, the sound produced by different instruments can be isolated via source separation [11], which could be deployed to zoom in on a particular instrument or instrumental section [12].

While the solutions mentioned above primarily rely on an analysis of the audio channel of the performance recording, the visual channel has remained underexploited. In addition to enabling the development of new functionalities of platforms like *RCO Editions* and *Berliner Philharmoniker's Digital Concert Hall* not covered by audio analysis, the analysis of the visual channel could also help to resolve some of the critical challenges faced by audio analysis. For instance, achieving reliable sound source separation is challenging in the case of large ensembles where the sound produced by many different instruments overlaps both in time and frequency [7].

In this paper, we focus on the analysis of the visual channel of the audio-visual recording of a symphonic music performance and address the problem of annotating the activity of individual musicians with respect to *whether they play their instruments at a given timestamp or not*. The envisioned output of the solution we propose in this paper is illustrated in Fig. 1, where playing and non-playing musicians are isolated as indicated by respectively the green and red rectangles.

Knowing the playing (P) and non-playing (NP) labels for each musician allows the annotations of an audio-visual recording to be enriched in a way that is complementary and supportive to audio-only analysis. For instance, repeats and solo parts could be detected also by analyzing the sequences of P/NP labels to allow novel non-linear browsing functionalities (e.g., skip to solo trumpets, skip to "tutti"). The problem of performance-to-score synchronization, which is typically addressed through audio-to-audio alignment [22], could also be approached in a multimodal fashion by combining state-of-the-art auditory features and P/NP labels [5].

Related methods operating on the visual channel typically deploy a standard classification paradigm and learn visual models for human actions [28,39]. The disadvantage of this approach in the

* Corresponding author.
E-mail addresses: A.Bazzica@tudelft.nl, alessio.bazzica@gmail.com (A. Bazzica), C.C.S.Liem@tudelft.nl (C.C.S. Liem), A.Hanjalic@tudelft.nl (A. Hanjalic).

[1] http://www.concertgebouworkest.nl/en/rco-editions/
[2] http://www.digitalconcerthall.com/

http://dx.doi.org/10.1016/j.cviu.2015.09.009

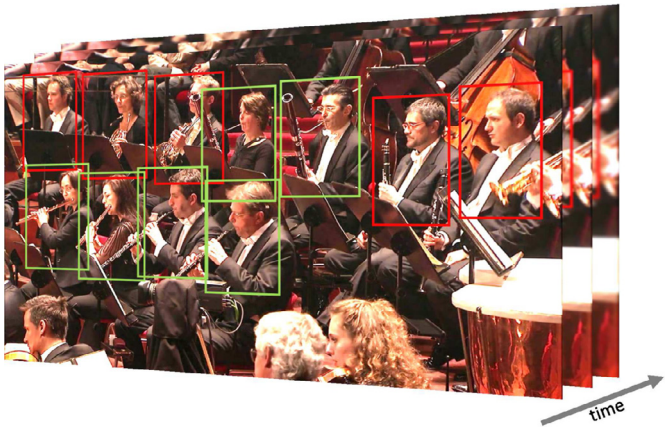


Fig. 1. Envisioned output of the method proposed in this paper. Green (red) bounding boxes mark the musicians that play (don't play) their instrument at a given time stamp. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

problem context of symphonic music concert videos is that the models may not be generic enough to cover the wide variety of instruments used and the ways the P/NP activities of individual musicians could be visually recorded. Additionally, a realistic view at the reliability of solving this classification problem reveals the need for manual human intervention in order to correct unavoidable classification errors, in particular in a professional context when high-quality annotation output is required.

The method we propose in this paper is geared not only towards neutralizing the disadvantage mentioned above, but also towards incorporating human intervention in the way that is as efficient and effective as possible. We implement our proposed solution to assign P/NP labels per musician to the timeline of a symphonic music performance as a modular framework so that we can provide answers to the following research questions:

- **RQ1**: How reliably can we isolate clusters of images depicting individual musicians from the keyframes extracted from a music video?
- **RQ2**: How accurately can sequences of P/NP labels be generated?
- **RQ3**: What is the tolerance of the proposed framework to errors in different modules?
- **RQ4**: Is a static image informative enough to reveal whether a musician is playing an instrument?
- **RQ5**: What is the relation between the amount of human intervention and the quality of the obtained P/NP label sequences?

The paper is organized as follows. We start by explaining in Section 2 the context in which we operate in this paper and that characterizes the realization and recording of a typical symphonic music performance. By taking into account the properties of the work context and the related limitations, we proceed in Section 3 by analyzing the usability of the existing related work and in Section 4 by stating our novel contribution and explaining the rationale behind our proposed framework. We introduce the notation, set the goals and make assumptions in Section 5. We present our method in Section 6 elaborating on the realization of different framework modules. After we explain the experimental setup in Section 7, we present our assessment of the framework in Section 8 where we also provide answers to the research questions posed above. We conclude with a discussion section in which we also present future research directions (Section 9).

2. Characteristics of a symphonic orchestral recording

A symphonic orchestra consists of a large number of musicians organized in *sections* (string, brass, woodwind or percussion). Sections are further divided into *instrumental parts*. Each instrumental

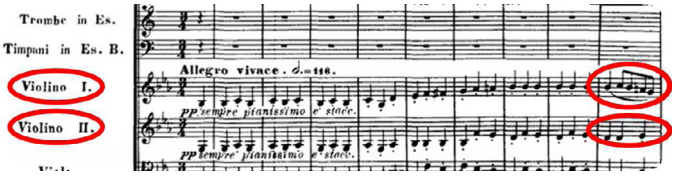


Fig. 2. Excerpt of a score: same instrument, different instrumental parts.

Fig. 3. Examples of video frames showing different settings of musicians and their instruments on the stage during the symphonic music performance.

part consists of a number of musicians playing one particular instrument and following a specific musical score. For instance, in Fig. 2 the instrumental parts "Violino I" and "Violino II" play different notes even if the instrument is the same (violin). According to the scores, when one musician belonging to one instrumental part is (not) playing, all the other musicians performing the same instrumental part are expected to be (in-)active as well. This usually holds even in the *divisi* case[3].

Performance recordings may differ depending on several factors like, for instance, the type of environment (indoor vs. outdoor), the number of cameras and whether camera motion occurs. In this paper we focus on the indoor case and we consider two possible types of recording: single- and multiple-camera recordings. The former is made from a fixed point of view and with a fixed zoom factor. In this way, the whole ensemble is always visible and each musician covers the same region of the video frames throughout the video. The latter typically involves multiple-cameras positioned around and on the stage, with the possibility to zoom and pan. This type of recording typically serves as input to a team of experts in order to create an edited video using a script (e.g., "when the 100th bar of the scores starts, the 3rd camera switches to a close-up on the first clarinet player"). Thereby, the visual channel mainly focuses on (parts of) the orchestra, but can also show the conductor and the audience in the concert hall.

Both in the single- and multiple-camera recordings, depending on the camera position, some musicians appear frontally, some non-frontally, and some even from the back, (fully) occluding their instruments. As illustrated in Fig. 3, the setting of the orchestra on the stage is rather dense, resulting in significant occlusion of individual musicians and their instruments. A video frame taken from the visual recording of the performance can therefore contain multiple musicians, not necessarily belonging to the same section or instrumental part.

The characteristics of the context in which we operate, as described above, have significant impact on the extent to which we can rely on the existing related work in conceptually developing our proposed solution, but also on the way how we approach the definition and implementation of the modules of our framework. This will be explained in more detail in the following sections.

3. Related work

The problem of extracting the sequence of P/NP labels for each musician continuously over time from an audio-visual recording of a

[3] http://en.wikipedia.org/wiki/Unison#Divisi

symphonic music performance has not been directly tackled so far. We explore here, however, the usability of a number of related approaches.

3.1. Detecting the playing/non-playing activity

Regarding the detection of P/NP activity in general, we classify the existing work into hardware-based, score-based, audio-based and vision-based approaches.

3.1.1. Approaches based on dedicated hardware

Probably the most intuitive approach to inferring the activity of a particular musician is via dedicated hardware [19,30]. While theoretically effective, the critical deficiency of such an approach is that it requires obtrusive settings, which are unnatural in the work context described in the previous section. For instance, a webcam may need to be mounted above the vibraphone in order to detect which bars are covered by the mallets [30].

3.1.2. Score-based approaches

An alternative to deploying obtrusive dedicated hardware is to rely on the data from the regular audio-visual recording, possibly in combination with the available supplementary material. For instance, the P/NP states could be inferred by analyzing a synchronized music score, that is, by looking at presence of notes and rests in each bar as done in [5]. Such a method allows to infer P/NP labels for every instrumental part at every time point. However, as pointed out in [11], even if full scores are freely available for many classical pieces, they are rarely aligned to a given audio recording. In order to pursue this strategy, the score and the performance recording need to be synchronized using existing alignment methods [14,24]. Performing such synchronization can be challenging, especially in presence of structural variations between the score and the recording (e.g., omission of repetitions, insertion of additional parts). However, in practice, even though partial alignment methods exist, likely failures in the structural analysis and subsequent segment matching steps can lead to corrupted synchronization results [20,31].

3.1.3. Audio-based approaches

Source separation techniques could be considered to isolate the sound of each instrument and infer P/NP labels by analyzing the isolated instrument-level signals. In view of the context in which we operate, however, this approach is not likely to be successful. Typically, only a limited number of instruments can be recognized with an acceptable accuracy. In [18], the authors address the challenging problem of recognizing musical instruments in multi-instrumental and polyphonic music. Only six timbre models are used, hence this approach has limited utility for symphonic orchestras where more models would be needed. In [4] the number of recognized instruments is 25, but the recognition is performed in those parts of a piece in which a single instrument is played alone. This limits the applicability of this approach in our work context to the rare solo segments only. While it was shown in [11] and [33] that effective audio source separation needs prior information derived by synchronized music scores, such an informed source separation approach would include the limitations of those related to score synchronization, as discussed above.

3.1.4. Vision-based approaches

Insufficient applicability of audio-based approaches in our work context makes us investigate the alternatives relying on the visual channel. When video recordings are available, we can see musicians interacting with their instruments. They hold them in a certain way when playing, while they assume different body poses when not playing. In the former case, musicians also move in order to make music (e.g., bowing, pressing keys, opening valves, moving torso to help blowing). Hence, visual appearance and motion information could be potentially useful in inferring whether musicians are playing or resting.

Fig. 4. Examples of the setting of musicians and their instruments as considered by the existing vision-based approaches.

In view of the above, one could explore human-object interaction (HOI) by analyzing visual object appearances in a static image – i.e., a keyframe extracted from a video. For this purpose, investigation of presence of objects of interest (in this case, music instruments), spatial relationships between objects and human body parts has been found promising [38,39]. Dedicated datasets have been developed for this line of research, a good example of which is the “people playing musical instrument” (PPMI) dataset [38].

Alternatively, in video action recognition, both visual appearance and motion information are exploited [23,25,28]. State-of-the-art performance with popular datasets, like the UCF101 [29], shows that several actions, like “playing violin”, can be detected.

The aforementioned methods for HOI detection and video action recognition are based on a supervised classification approach. While such methods are sophisticated and in general have the potential to outperform previously discussed non-visual approaches, they require visual input of a particular type in order to train reliable classifiers. For example, as illustrated in Fig. 4, the PPMI dataset consists of images containing sufficiently large and well visible regions corresponding to a human and an instrument. This makes the aforementioned methods not applicable to the situations addressed in this paper and illustrated by the orchestra settings in Fig. 3.

3.2. Detecting, isolating and recognizing musicians

In order to design a system which yields a sequence of P/NP labels for each musician, we first have to solve the *musician diarization* problem. In other words, we want to understand which musician appears when and where in the video frames. The related literature for this task includes works about detecting, tracking and recognizing people in videos. Then, for each musician appearing in the scene, the regions of the video frames which are informative for the inference of the sought P/NP labels have to be isolated by means of image segmentation.

When the input video consists of a set of fixed-camera recordings, the positions of the musicians in the scene can be manually annotated using a reference video frame from each video (e.g., the first one). Such a manual initialization step is time inexpensive and can be done because the musicians do not change their position throughout the performance. Therefore, the annotated coordinates can be used for the whole recording.

In the case of a video recording consisting of different shots resulting from camera zoom-in and pan actions, manual-only annotation of musicians becomes too complex and needs to be helped by automatic visual analysis tools. Off-the-shelf face detectors, face clustering and recognition methods can be deployed for this purpose, possibly supported by a face tracking algorithm to collect and verify the evidence from consecutive video frames [9,27].

Specifically related to face clustering, state-of-the-art solutions are typically based on context-assisted and constrained clustering [37,40], possibly including human intervention in order to produce high quality results [41]. For instance, clothing information is exploited to discriminate people with similar faces but dressed differently [40]. *Cannot-link* constraints are used to avoid that two faces detected in the same image fall into the same cluster. People can be tracked and *must-link* constraints can be inferred by the generated face tracks [37]. Face-related visual features can be extracted for every detection, or just when the estimated quality of the face image is good enough to extract reliable information [3]. Finally, to avoid that too many face clusters are generated for the same identity, semi-automatic algorithms can be used to iteratively merge clusters [41].

The existing methods are typically tested only on frontal faces. Alternatively, as done in [3], the detected profile faces are continuously tracked over time, but used at the clustering step only when a switch to a (near-)frontal view occurs. In view of our problem context described in Section 2, this focus on (near-)frontal faces makes the methods described above insufficiently suitable as modules of our envisioned framework. This was also revealed by an initial investigation we performed to inform the design choices for this framework, the results of which are reported in Section 7.1.1.

4. Contribution and rationale

In view of the fact that the visual channel of the symphonic music recordings is available, and based on the conclusions drawn in Section 3.1 regarding the performance-related and practical disadvantages of hardware-, score- and audio-based methods, in our approach we focus on the visual channel to infer the P/NP activity per musician. In order to cope with inevitable errors of automated visual analysis of challenging HOI cases in our application context and to secure high accuracy of the obtained P/NP label sequences, we choose for a semi-automatic approach, where human intervention is efficiently and effectively combined with automated analysis. The value of such hybrid approach for video annotation has already been shown in the past (e.g., [35]).

The proposed method involves two main steps, musician diarization and label assignment per musician and time stamp. Learning from the analysis of the related work, we pursue the development of the solutions for both steps by making the following critical design choices.

Regarding the musician diarization step, as argued in Section 3.2, we need a more reliable method for identifying the musicians than what the state-of- the-art in the field currently offers. While we can rely on standard face detection methods, the choice of the face clustering method leaves room for improvement, primarily in view of the requirement to obtain the face clusters that are as pure as possible. This purity is essential because errors in clustering directly propagate to the resulting P/NP label sequences. We have initially considered the approach described in [41], which semi-automatically merges an initial set of face clusters assuming that all of them are close to being 100% pure. However, our preliminary experiments deploying this method on our concert video data have revealed that only a part of the generated clusters can be obtained as almost 100% pure, while the remaining clusters are too noisy. Moreover, as reported in Section 7.1.1, we found that different features and image regions from those reported in [41] may yield much better face clusters on our data. We therefore investigated alternative ways to increase the number of pure face clusters by strategically employing human annotators. Beside alleviating the impact of unavoidable non-pure clusters, such a semi-automatic strategy can be exploited to efficiently and effectively reject clusters of non-relevant targets – i.e., conductor and audience but also false face detections. Our approach turns out to require significantly simpler visual analysis tools than the complex, sophisticated person identification methods discussed in Section 3.2.

Once the musician diarization problem is solved, we infer the P/NP activity per musician. As opposed to the methods discussed in Section 3.1.4, we deploy the information in the visual channel in such a way to better exploit and match the characteristics of the work context we address, however, at the same time, being able to handle the full scope of content generated in such context – i.e., any performance of any symphonic orchestra. Specifically, instead of aiming to develop generic HOI models via a supervised learning approach, we base our solution on the clustering principle. We search for clusters ad-hoc, for a given video of a performance. Thereby, we focus on the following cluster categories in which we group the detected musicians' images: (i) musician identity, (ii) point of view and (iii) playing/non-playing activity. Creating clusters for these categories, labeling them appropriately and propagating the labels to the individual video frames will then automatically result in the targeted P/NP label sequences. The advantage of this approach, as opposed to those based on training the HOI models, is that there is no dependence on the type of instruments nor on the actual way how HOI is represented – i.e., in which way a musician interacting with her instrument is depicted in a particular recording, as long as HOI activity is depicted consistently along the video. In our work context, consistency in general can be assumed due to the following characteristics: (a) the number of musicians is limited, (b) the setting of the orchestra on the stage is constant within one performance, and (c) the variations by which musicians appear in the video are limited by the limited number of camera views.

In view of the above, our proposed approach can now conceptually be summarized as follows. By exploiting the redundancy of each analyzed video recording (e.g., multiple occurrences of the same camera angle), we accumulate information on the dominant appearances of various musicians in terms of their instrument-playing activities. These dominant appearances are then turned into clusters that coincide with P/NP activities to be labeled accordingly, through human intervention. This combination aims to achieve high level of output quality eliminating the need for extensive model training and making the annotation problem more tractable. We refer to Section 6 for a detailed explanation of the different steps of our method.

5. Notation, goals and assumptions

Given a multi-camera video recording of a symphonic music performance, we aim at inferring for each performing musician the P/NP labels over time. A label is assigned at regular intervals (e.g., every second) at the time point t starting from the first frame in the video. The videos generated by different cameras are denoted as the set $V = \{\mathbf{v}_i(n)\}$ where $n \in \mathcal{N} = \{0 \dots L-1\}$ denotes the frame index and L is the total number of frames. All the videos are synchronized in time and have the same length L. We further denote by M_{GT} the set of performing musicians, where GT stands for "ground truth", and by $|M_{\mathrm{GT}}|$ the set size.

In view of this notation, our goal is to learn the function $\mathrm{PNP}_m(t) : \mathcal{T} \rightarrow \{\mathrm{P}, \mathrm{NP}, \mathrm{X}\}$, which determines the P/NP label at the time points $t \in \mathcal{T}$ for each musician $m \in M_{\mathrm{GT}}$. The additional label X represents the cases when the label is not determined. As discussed in Section 7, we evaluate the accuracy of the learned PNP functions as well as the amount of determined P/NP labels.

While we count on multi-camera recordings (see Section 2), the minimum required number of cameras for our approach is one and camera motion is allowed (e.g., panning, zoom-in, zoom-out). The method does not require information about which instruments are played during the performance. Furthermore, we do not make any assumption regarding the *timeline coverage*. In other words, while we do not require that every musician is continuously captured by a camera during the performance, it can also happen that at a given time point the same musician is captured by multiple cameras. We only require that for each musician $m \in M_{\mathrm{GT}}$ her corresponding

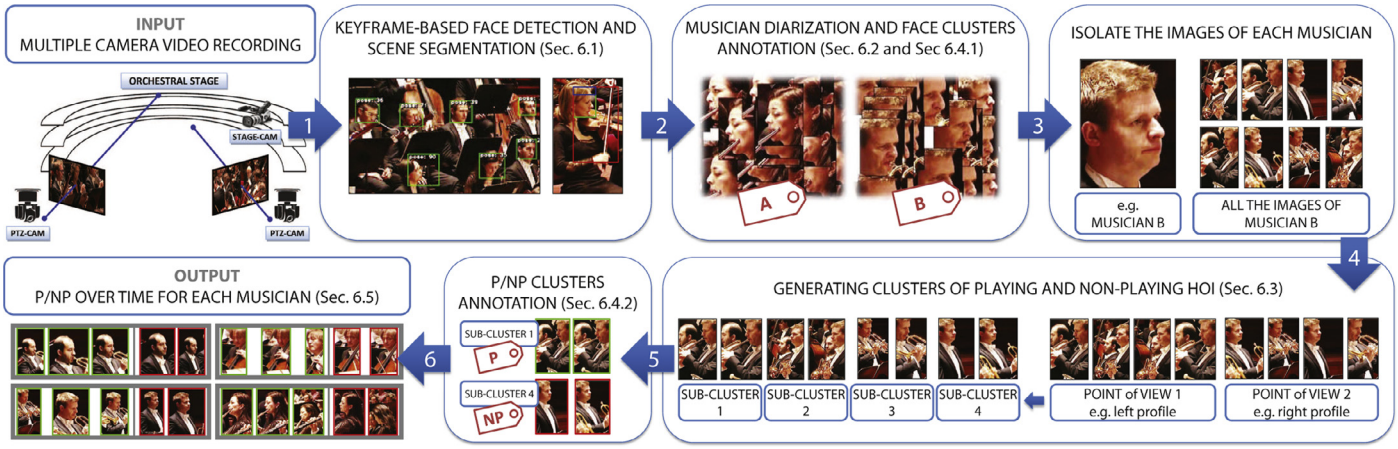


Fig. 5. Illustration of the modular framework implementing the proposed method for extracting P/NP labels per musician from a video recording.

instrumental part is known. This knowledge allows to make a partition of $M_{GT} = \cup_{h=1}^{H} M_{GT}^{h}$ into H mutually disjoint subsets and to recover part of the missing P/NP labels as described in Section 6.6.

6. Method description

In this section, we describe the framework representing the realization of our proposed method and illustrated in Fig. 5. First, the keyframes are extracted from the given multiple-camera recording and processed to detect and isolate musicians in the scene (details reported in Section 6.1). A musician diarization problem is then solved by combining face clustering and human annotation (respectively discussed in Sections 6.2 and 6.4.1). In this way, all the images belonging to each performing musician will be effectively and efficiently isolated and linked to the correspondent musician identity label.

At this point, instead of using pre-trained visual models which independently infer playing and non-playing labels for each single image, we rely on a novel unsupervised method for the reasons discussed in Section 3.1.4. Such method, described in Section 6.3, aims at learning ad-hoc discriminative visual patterns for each performing musician to be used for distinguishing playing activities from non-playing ones. This approach produces sub-clusters of P/NP images which will be manually labeled accordingly using the procedure described in Section 6.4.2. Finally, the sought P/NP label sequences are computed as described in Section 6.5.

6.1. Keyframe-based face detection and scene segmentation

For every video $\mathbf{v}_i(n) \in V$, one keyframe $\mathbf{f}_i^k$ is extracted at predetermined time points n_i^k (e.g., at regular intervals) where n_i^k is the kth time point for the ith video. The set of keyframes extracted from $\mathbf{v}_i(n)$ is denoted as $F_i = \{\mathbf{f}_i^k\}_{k=0}^{K-1}$.

For each keyframe, we detect faces and estimate the head pose angle. Regarding face detection, we rely on standard, off-the-shelf approaches, as described in detail in Section 7.3.1. In this way we build the sets $D_i^k = \{\mathbf{d}_i^{k,l}\}$, where $\mathbf{d}_i^{k,l}$ is the lth detection in the keyframe $\mathbf{f}_i^k$. Each detection $\mathbf{d}$ is defined as $(\mathbf{b}, \theta)$ where $\mathbf{b} = (x, y, w, h)$ is the vector encoding the face bounding box geometry and $\theta \in [-90, +90]$ is the estimated head pose.

Finally, we exploit the face bounding box geometry using simple but effective heuristics to identify visual information supplementary to the face that can be valuable for the subsequent clustering steps. Here we focus in particular on the hair and upper body of the musician, and related to the latter, on those regions where the instrument can be expected. Given a face detection $\mathbf{d} = (\mathbf{b}, \theta)$, the two additional bounding boxes are inferred using the Vitruvian man ratios as done in [9]. The hair bounding box is defined as $(x, y - h/4, w, h/4)$. As for the upper body segmentation, we extend the heuristic presented in [9], which is limited to the frontal faces, in order to infer the region of interest for any value of $\theta \in [-90, +90]$. The upper body bounding box is therefore computed as a function of $\mathbf{b}$ and θ. The underlying idea is to look at the region of the image in the direction of the musician's gaze where we expect to see the instrument. If $\theta > \theta^\star (< -\theta^\star)$, we look at the right(left) side of the face bounding box. When $\theta \in [-\theta^\star, +\theta^\star]$, we center the face bounding box horizontally. $\theta^\star$ is the critical angle used to discriminate frontal and profile faces. The upper body region includes the head and the torso. The torso has a height of $2.6 \times h$ and a width of $2.3 \times w$ [9]. An illustration of the results of this segmentation process is given in Fig. 6.

6.2. Musician diarization via face clustering

Grouping the detected faces into clusters of individual musicians can be performed in different ways. We consider four possibilities that we refer to as (i) *unconstrained*, (ii) *context-assisted*, (iii) *constrained*, and (iv) *context-assisted and constrained*. The *unconstrained* method relies on the visual information only consisting of visual features extracted from the face and hair regions. In addition to visual information, *context-assisted* methods also rely on the *visual context* of the detected face. The upper body region extracted for a face may help discriminating between those musicians whose faces look similar, but who play different instruments. Similarly, a scene descriptor could be deployed to discriminate between similar faces belonging to musicians placed in different parts of the orchestra. In the *constrained* method, we again deploy face- and hair-related visual features, but also exploit the fact that multiple face detections in the same frame should belong to different identities. We build a sparse matrix of cannot-link constraints CL for each pair of faces $(\mathbf{d}_i^{k,l}, \mathbf{d}_i^{k,l'}) | l \neq l'$ detected in the same keyframe. CL is then used to ensure that multiple detections in the same keyframe fall in different face clusters. Another type of constraint which could be deployed is the must-link constraint. During a shot, the detected faces could namely be tracked and therefore linked over time. However, taking this into account would increase the complexity of the system and might not generate exact constraints as in the case of the CL set (e.g., due to the mistakes with crossing face tracks generating wrong must-link constraints). Finally, the *context-assisted and constrained* method exploits both visual context information and the cannot-link constraints.

As for choosing a suitable number of clusters, we consider the following information that can be reasonably defined a priori. The number of musicians $|M_{GT}|$ may vary, but a typical symphonic ensemble ranges from 50 to 100 players. In addition to the orchestra, some

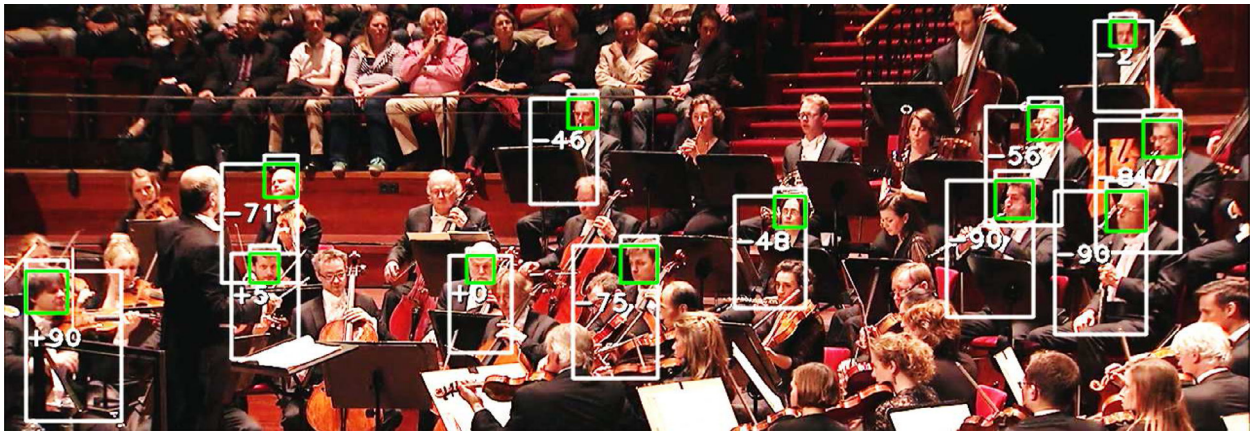


Fig. 6. Example of keyframe segmentation. For each detected face, the upper body region is determined considering the estimated head pose. In this way, we find the region of the image where the HOI is expected to be visible.

of the frames also show the conductor and the audience. Together, the musicians, conductor and audience form the set E of "entities" to be isolated. Furthermore, the same entity can be recorded from different cameras/viewpoints, and also with variations (e.g., due to camera zoom-in). Therefore, the number of expected clusters can be estimated as $\lceil \alpha \times |E| \rceil$, where $\alpha \geq 1$ is a factor which accounts for the number of cameras and additional variations in the types of the recorded visual material.

The values for α and $|E|$ can be chosen rather freely, as long as they are large enough. This is due to the subsequent labeling step in which all the detected clusters where musician m appears are merged together into one set S_m containing all the detections $\mathbf{d}_i^{k,l}$ of that musician, independent of the camera viewpoint, HOI activity or other variations. Therefore, while the detected clusters should be sufficiently pure, over-segmentation is not problematic. The labeling step is performed manually and is explained in Section 6.4.1.

6.3. Generating clusters of playing and non-playing HOI

Once the set S_m is generated, we follow the hypothesis that the images contained in there can be distinguished from each other using two dominant dimensions: camera viewpoint and performed HOI action. Under this assumption, we divide each set S_m into sub-clusters. Each sub-cluster should contain the images of the musician m with one specific HOI action recorded from a specific camera viewpoint. This results in a set of C_m mutually disjoint subsets S_m^c such that $S_m = \cup_{c=1}^{C_m} S_m^c$. We estimate the number of sub-clusters C_m by first estimating the number of camera viewpoints $|\text{PoV}_m|$ on the musician m. Then, the number of sub-clusters corresponding to a playing or non-playing HOI is $2 \times |\text{PoV}_m|$.

The number of viewpoints on a musician m is estimated as follows. To maximize the accuracy of the clustering process at this stage, compared to Section 6.2, we choose for a more sophisticated method for estimating $|\text{PoV}_m|$. We do this by analyzing how the bounding box geometry $\mathbf{b}$, the head pose θ and the camera/video index i values are distributed. By empirical evaluation, we found that the number of dense regions formed by the set of $(w \times h, i)$ pairs, respectively the face bounding box area and the camera/video index of each detected face $\mathbf{d}_i^{k,l}$ belonging to m, is a suitable and consistent choice.

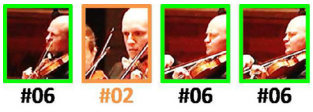
Then, in order to generate the sub-clusters S_m^c, we follow these steps:

1. for each $\mathbf{d}_i^{k,l} \in S_m$, we extract an image $\mathbf{I}_i^{k,l}$ from the keyframe $\mathbf{f}_i^k$
2. for each image $\mathbf{I}_i^{k,l}$, we extract a vector $\mathbf{x}_i^{k,l}$ of visual appearance features
3. we build a descriptor matrix $\mathbf{X}_m$ having $|S_m|$ rows, where each row is a feature vector $\mathbf{x}_i^{k,l}$
4. we cluster the detections in S_m by running a clustering algorithm taking $\mathbf{X}_m$ as input and with the number of clusters to be generated being set to C_m.

Fig. 7. Example of labeled sub-clusters generated for a flute player (only a few representative images per cluster are shown).

In order to assess the informativeness of different regions of the image, we consider two options for extracting $\mathbf{I}_i^{k,l}$, which capture the face and the upper body regions. As for the way we visually describe the segmented images, we consider global and local features. As for the latter, we aim at exploiting as much as possible the redundancy of the images belonging to each musician. We therefore train one visual word vocabulary for each set S_m instead of training a vocabulary for the whole recording. By training ad hoc vocabularies, we expect that the discriminative power of the trained visual words is optimized for each musician. In Section 7.1.2, we report the details about the used features and the optimal parameters (e.g., number of visual words).

The obtained sub-clusters directly imply the P and NP labels to be assigned to them and therefore the quality of sub-clusters also determines the quality of our P/NP annotation framework. We explain the sub-cluster annotation process in Section 6.4.2. Examples of labeled sub-clusters are shown in Fig. 7. Unlike in the case of face clusters labeling, non-pure or otherwise ambiguous sub-clusters are not discarded, but annotated using the label X (undetermined).



(a) Type A1: the dominant musician in S_6^c is #06, the dominant label is P and the image of #02 is also P.

(b) Type A2: the dominant musician in S_6^c is #06, the dominant label is P but the image of #02 is NP.

(c) Type B: the sub-cluster S_{36}^c contains only images of #36, however it is non 100% P/NP pure.

Fig. 8. Examples of different types of error generated at the face and/or the PNP clustering steps. While the errors of type B have a direct negative impact on the accuracy, an error of type A1 or A2 leads to a P/NP labeling error depending on the timestamp of the detection belonging to the "wrong" musician.

This clustering step is fundamental to make the subsequent human annotation process efficient. In fact, if every single detection were manually annotated, the complexity of the human annotation task would be $O(|M_{GT}| \times L)$ – i.e., linear to the number of musicians multiplied with the temporal length of the recording. Since we assumed that the number of points of view $|\text{PoV}_m|$ is limited, the complexity of the human annotation task using our approach becomes $O(|M_{GT}|)$ – i.e., linear to the number of musicians.

6.4. Human annotation

Our proposed framework illustrated in Fig. 5 involves two manual labeling steps, the first one annotating the face clusters by the corresponding musician ID and the second one annotating the sub-clusters in terms of P and NP labels.

In general, the annotation process of a cluster of images works as follows. The annotator inspects the content of a given cluster which is rendered, for instance, as a grid of images. Then, the *purity* of the given cluster is evaluated. A cluster is pure if most of the images belong to one class. We call such class *dominant*. If there is a dominant class, it is used as label for the cluster. Conversely, a non-pure cluster is discarded in order to prevent that the labeling accuracy will be low. We assume that: (i) human annotators are able to detect the presence of a dominant class, and (ii) human annotators can recognize the dominant class (if present). More details about the two manual labeling steps are reported below.

6.4.1. Face clusters annotation

The annotator is provided with a reference table of musician IDs like the one in Fig. 9. The images within a face cluster are shown to the annotator and the annotator decides first whether the cluster is pure enough, that is whether the cluster has a *dominant* musician ID.

If the annotator finds the cluster to be pure enough, then she uses the reference table to check whether the dominant identity belongs to one of the musicians. If a musician is dominant in the face cluster, then the corresponding label is chosen and automatically propagated to all the face detections belonging to the given cluster. In the cases of a non-musician dominant label (conductor, audience or non-face images) and a non-pure cluster, the cluster is discarded and the face detections belonging to it will not be used anymore.

A first type of error that can occur at this step is the error of *type A* (e.g., Fig. 8 a and b): if a cluster is not discarded and therefore labeled with $m \in M_{GT}$, any image not belonging to the musician m will generate a musician labeling error. The impact of this error type on the accuracy of P/NP labeling is discussed in more detail below.

6.4.2. P/NP clusters annotation

For this task, the annotator does not need any reference table and we expect that no specific expertise is required in order to distinguish playing and non-playing actions for any musical instrument. We also assume that each sub-cluster can be annotated independently.

Given a sub-cluster S_m^c to be labeled, the annotator once again decides first whether it is sufficiently pure. Differently from the previous annotation task, the purity now has *two* dimensions. The first one is related to the presence of a dominant P/NP class, that is whether the majority of the images show either a playing or non-playing HOI. When a dominant class is chosen, all the images not belonging to that class will generate a P/NP labeling error of *type B* (e.g., Fig. 8c). The second purity dimension deals with the error of type A since a sub-cluster may contain images of other musicians due to errors at the face clustering phase. Considering these two aspects, we assume that a sub-cluster is discarded if it contains too many errors of type A and/or B.

Fig. 9. Example of reference table provided to the face clusters' annotators.

Finally, regarding the error of type A, we distinguish two cases occurring when a P/NP cluster S_m^c is not discarded and contains images belonging to one or more musicians $m' \neq m$. The error of *type A1* occurs when an image of a different musician m' has the same P/NP label as the one which is dominant in the sub-cluster (e.g., Fig. 8a). The error of *type A2* occurs when an image of a different musician m' has not the sub-cluster's dominant P/NP label (e.g., Fig. 8b). The main impact of these types of error is that a spurious observation is added to the musician m and removed from the musician m'. Then, for the musician m, the system may generate an additional and eventually wrong P/NP label according to factors, which depend on the way P/NP sequence are generated as explained in Section 6.5.

6.5. Generating sequences of P/NP labels

Taking the sub-clusters labeled as either P, NP or X and the keyframe's timestamps associated to the images belonging to the sub-clusters as input, we now proceed with generating the function $\text{PNP}_m(t) : \mathcal{T} \rightarrow \{\text{P}, \text{NP}, \text{X}\}$ that produces the P/NP/X label sequence for each musician $m \in M_{GT}$.

As defined in Section 5, we aim at reconstructing the PNP sequence for every musician at regular time intervals (e.g., every second). The reason why we do not extract the labels for every frame lies in the inherent nature of the P/NP labels. As explained in [5], it is not likely that two or more P/NP switches occur in a short period of time, because during short musical rests musicians keep a playing body pose. Hence, we adopt the same sliding window approach of [5] and we derive P/NP labels periodically for every musician. A large

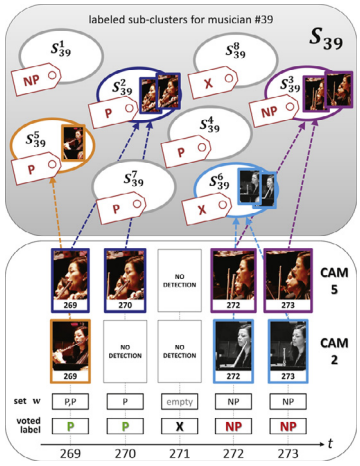


Fig. 10. Illustration of the process of creating the P/NP label sequences for a musician m via majority voting. In this example, we focus on the case of two cameras recording the musician $m = 39$ and we set the sliding window size to 1 second (for the sake of simplicity). Given 8 labeled sub-clusters S^c_{39}, every second the available P/NP labels are sought in the labeled sub-clusters. The retrieved labels are used to build the sets $\mathbf{w}$ to which a majority voting is applied to determine the final label.

window size (e.g., 5 seconds) accounts for the time required to switch from a playing to a non-playing body pose (and vice versa).

For each musician m, each label is generated through a voting process illustrated in Fig. 10. At every timestamp $t \in \mathcal{T}$, a set $\mathbf{w}$ is built by exploiting the labeled sub-clusters S^c_m associated with the musician m as follows. We look for the images $\mathbf{I}^{k,l}_i \in S^c_m$ extracted within the current sliding window time interval. This search can lead to a variable number of results, depending on how many cameras record m in the considered period of time. For each found image, one P/NP label is added to $\mathbf{w}$, inherited from the sub-cluster the image belongs to. Discarded sub-clusters are ignored. Consequently, $\mathbf{w}$ is either an empty set or contains one or more labels. In the former case, the label assigned at the timestamp t is X because there is no observation of m in the considered time window. In the latter case, P(NP) is assigned if the number of P(NP) labels in $\mathbf{w}$ is greater than the number of NP(P) labels. If the numbers of P and NP labels in $\mathbf{w}$ are equal, the label X is assigned.

6.6. Dealing with missing observations

As pointed out in Section 5, there is no guarantee that each musician is always visible from at least one camera. If a musician does not appear in a keyframe, no P/NP label can be inferred using the procedure explained above. However, the domain knowledge on the orchestral setting (Section 2) allows us to infer the labels for individual musicians from all the other musicians playing the same instrumental part and thus belonging to the subset M^h_{GT}. In this case, for each subset M^h_{GT}, the expected sequence of labels is the same for every musician $m \in M^h_{\mathrm{GT}}$.

We propose two different strategies to extrapolate the labels: (i) *highest timeline coverage* (highest TC), and (ii) *merging*. Given M^h_{GT}, the highest TC approach assigns one of the existing PNP functions to all other musicians in M^h_{GT}. The optimal PNP function for a given instrumental part h is that computed for the musician $m^\star$ such that $m^\star = \arg\min_{m \in M^h_{\mathrm{GT}}} |\{t : \mathrm{PNP}_m(t) = \mathrm{X}\}|$. The rationale behind this strategy is to base the extrapolation on the musician for which the number of observations is maximized. Differently, the merging strategy computes a new PNP function for each instrumental part by combining all the labeled sub-clusters S^c_m belonging to the musicians performing the considered instrumental part. As opposed to relying on the strongest evidence as in the previous strategy, here we combine all the available evidence belonging to a certain instrumental part. For this purpose, we deploy a modified version of the majority voting approach described in Section 6.5. When $\mathbf{w}$ is populated, instead of considering the sub-clusters S^c_m of a single musician, we consider *all* the sub-clusters S^c_m such that $m \in M^h_{\mathrm{GT}}$.

7. Experimental setup

In this section, we detail how we implemented the proposed framework, present our dataset, and explain how we conducted the experimental evaluation.

7.1. Framework implementation

The design choices and the parameter selection underlying the realization of our framework (presented in Section 6) were informed following the protocol described in Sections 7.1.1 and 7.1.2.

7.1.1. Musicians diarization

We describe the way we implemented the four face clustering methods introduced in Section 6.2 and explain how we selected features and parameters.

The *B-cubed precision/recall* [1] was adopted to assess the quality of the produced clusters. We chose the number of face clusters by approximating the number of entities $|E|$ to the number of musicians. In the case of the development set, $|E|$ was set to 7. A suitable value for factor α taking into account the variations of various types was found by inspecting multiple options, namely 1, 1.5, 2, 2.5, 3, 4, 5, 10, 15 and 20 (generating from 7 to 140 face clusters).

For clustering itself, we used k-means in the unconstrained case and COP k-means [36] in the constrained one. The constrained face clustering methods were not assessed using the development set because COP k-means has no parameters to be tuned and the number of cannot-link constraints generated for the development set was too low.

As for the unconstrained face clustering, we considered two options, both relying on state-of-the-art visual features. In the first one, we deployed Local Binary Patterns (LBP) and Histograms of Oriented Gradients (HOG) from the face bounding box as done in [41]. In the second one, we focused on the face bounding box extended to the hair region from which we extracted Pyramid HOG (PHOG), Joint Composite Descriptor (JCD), Gabor texture (Gabor), Edge Histogram (EdgeHist) and Auto Color Correlogram (ACC) [17]. In both cases, we evaluated the impact of applying the Principal Component Analysis (PCA) [15] retaining 99% of the total variance.

In the context-assisted case, we included a description of the scene and/or a description of the upper body region. As for the former, given a detection $\mathbf{d}^{k,l}_i$, we extracted the JCD, PHOG and ACC global features from a downsampled copy of the keyframe $\mathbf{f}^k_i$. The upper body region was always described by LBP, PHOG, JCD, Gabor, EdgeHist and ACC. For both scene and upper body descriptors, we assessed the impact of including and excluding this information and also the option of including it by first applying PCA retaining a number of possible ratios of total variance (namely, 50, 70 and 99%).

By inspecting the results summarized in Fig. 12, we found that the optimal set of features to assess the face similarity is that extracted from the face-hair region and consisting of PHOG, JCD, Gabor, EdgeHist and ACC applying the PCA (see Fig. 12a). By comparing the plots in Fig. 12, we see how different combinations of contextual features affect the performance. The upper body features leads to the strongest improvement and the optimal ratio of retained variance for the PCA is 99% (see Fig. 12c). The scene features, whose optimal ratio of retained variance for the PCA is 70%, do not add a significant contribution (see Fig. 12b and d). Finally, the optimal value of α we chose was 15 because by increasing it to 20 we observe a saturation in the performance.

7.1.2. P/NP clustering

For each set S_m of images belonging to one musician, we estimated the number of points of view (see Section 6.3) as follows. The list of $(w \times h, i)$ pairs derived from S_m was first normalized (zero mean, unit variance). Then, we used DBSCAN [10] to automatically estimate the number of formed dense regions. We required that a dense region had at least 10 samples and the dense region radius parameter ϵ was set to 0.4. Pairs not belonging to any dense region were ignored.

As discussed in Section 6.3, the P/NP clusters S_m^c were produced considering two possible image regions and two possible types of feature. Evaluating on the dedicated development set, we found the following optimized global feature sets: face images were best described using Gabor, JCD and PHOG without applying the PCA, while upper body images by EdgeHist, Gabor, PHOG and ACC retaining 95% of the total variance. As for the local features, we considered two possible options, namely SIFT and OpponentSIFT [32], aggregating them either via bag-of-words (BoW) [6] or via spatial pyramid (SP) [16]. We also evaluated different visual words vocabulary sizes, namely 200, 400 and 1000 visual words (1000 only used with BoW). For each musician, that is for each set S_m, the visual word vectors were assigned via mini-batch k-means [26] applied to the visual words vocabulary training set, built by randomly sampling 500,000 feature vectors from the images in S_m. Using the development set, we found that the optimal way of describing both face and upper body images was using OpponentSIFT with 200 visual words, but aggregating the former via SP and the latter via BoW.

Image clustering was performed using the k-means algorithm. In order to assess the significance of the obtained results, we also included a random baseline method which simply works by randomly assigning the images in a given set S_m to the sub-clusters S_m^c.

7.2. Simulating the human annotation

In this work we address a number of research questions for which the experiment has to be repeated several times using different (types of) features and parameters. This is particularly true for the research question RQ3, for which we want to assess the overall impact of errors in different modules. In this context, deploying the two human annotation task presented in Sections 6.4.1 and 6.4.2 for every run is not feasible. In fact, in the full experiment we generate dozens of thousands of image clusters to be annotated. Another reason for not performing human annotation at this stage is that we do not know yet how to instruct human annotators with respect to how tolerant or strict they should be when coming across non-pure image clusters. We therefore made a number of assumptions and simulated human annotation using the available ground-truth information, also quantifying the perceived purity of a cluster of images and assessing the impact of different levels of strictness.

7.2.1. Modeling the human annotator

Following the annotation process and the assumptions reported in Section 6.4, we modeled a human annotator as follows. The core idea is to define a *rejection threshold* with which a cluster is discarded if the frequency of the dominant class is below such threshold. For each cluster, we compute a histogram of frequencies having one bin per class. If the highest frequency is below the rejection threshold, the cluster is discarded, otherwise it is kept and labeled with the dominant label. In our experiments, we used a number of distinct threshold values in order to study the impact on the overall performance. A high threshold corresponds to a *strict* annotator (high precision), while a lower value is a more *tolerant* one (balanced precision and recall).

7.2.2. Simulating the face clusters annotation

When labeling face clusters, we assigned the histogram bins as follows: one for each musician $m \in M_{GT}$, one for the conductor, one for the audience, and one for false positive face detections. We considered three types of human annotators by using the values 50, 70 and 90% for the rejection threshold. When the voted label did not belong to a musician, the face cluster was discarded. In order to understand to what extent face clustering is a critical step, we also used the face clustering ground-truth labels (*ideal* case).

7.2.3. Simulating the P/NP clusters annotation

When labeling a sub-cluster S_m^c, we computed the histograms assigning three bins associated to playing, non-playing and outlier images. The latter was used when an image of a different musician occurred, that is when an image belonged to a musician $m' \neq m$. We tested the following rejection thresholds: 50, 60, 70, 80 and 90%.

7.3. Dataset

We experimented on a dataset which in total consists of 29 videos (about 7 hours) from which we extracted more than 100,000 detections belonging to 105 different musicians. The dataset was built based on video recordings of two symphonic music concerts performed by two different professional orchestras and is representative for the context in which we operate, as described in Section 2. The first recording contains the four movements of Beethoven's Symphony No. 3 Op. 55, performed by the Royal Concertgebouw Orchestra (Amsterdam, The Netherlands) and it is a multiple-camera recording. The second one is a fixed, single-camera recording of the fourth movement of Beethoven's Ninth Symphony performed by the Simfònica del Vallès Orchestra (Barcelona, Spain). The two recordings, respectively referred to as "RCO" and "OSV", are available on request. To the best of our knowledge, there is no other available dataset consisting of real world data that we could have used alternatively.

7.3.1. RCO dataset

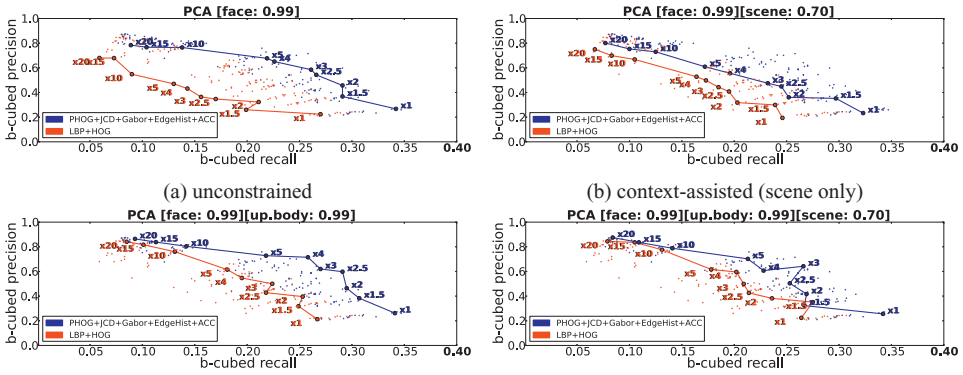
The RCO dataset (Fig. 11a) is organized into 4 sets of 7 synchronized videos where each set represents the multiple-camera recording of a movement (6 h and 40 min in total). The number of performing musicians is 54 and they are organized into 19 instrumental parts and playing 11 different instruments. The recording also captures the audience and the conductor. From each video, we extracted 1 keyframe every second producing 24,234 keyframes in total.

For each keyframe we detected the faces and estimated the head poses. This was done by combining a number of off-the-shelf multi-pose face detectors [34,42] via non-maximum suppression (NMS). The way we estimated the head pose is an adaptation of the method described in [3]. The adaptation was required in order to integrate the detector from [42] for which we initialized the confidence of its output to the acceptance threshold level (see [3]) in order to maximize the face detection recall. The choice of combining different types of detectors has significantly increased the number of detected faces. Overall, 66,380 face have been found which are distributed as follows: 1716 belonging to the conductor, 4539 to the audience, 3844 are false positives and the remaining 56,281 are distributed across the 54 musicians.

7.3.2. OSV dataset

The OSV dataset is a fixed, single-camera recording in which the performers appear at the same position throughout the whole event (see Fig. 11b). Faces approximately cover an area of 20×20 pixels, much smaller compared to those of the RCO dataset. The positions of the faces were manually annotated using a random frame as reference and then the head poses were, again manually, assigned to every face. Therefore, the face clustering step is not necessary for this recording since the musician identity is only a function of the face bounding box position in the reference keyframe. In this case, we extracted a keyframe every 2 s because, being the recording a fixed-camera one, oversampling in time would have been unnecessary for the goals of our experiment.

Fig. 11. Proposed datasets used in this work.



(a) unconstrained (b) context-assisted (scene only)

(c) context-assisted (upper body only) (d) context-assisted (scene and upper body)

Fig. 12. Face clustering evaluation on the dedicated development set. Each dot represented an evaluated combination of types of feature, amounts of retained variance while applying the PCA and factors affecting the number of generated face clusters. The lines are used to highlight the chosen combinations and how the performance changes when the number of generated cluster is changed (e.g., 10x means 10 times the number of musicians in the development set, namely 10×7).

7.3.3. Development set

As shown in Fig. 11a, part of the data extracted from the RCO dataset was used as development set. The reason why we did not include data from the OSV dataset there is twofold. First, we wanted to assess the general applicability of our method to an unseen recording. Hence, we followed a leave-one-recording-out approach while searching for visual features and parameters. Second, we find the RCO concert a more general case than the OSV due to the additional variations caused by panning and zoom-in camera actions.

The face clustering development set was generated by randomly sampling 1575 face detections belonging to the conductor, audience, 7 musicians performing different instrumental parts and also belonging to the false detections.

The development set was used to inform the design choices and select parameters of our framework. All the remaining data was used at the evaluation step.

7.4. Ground truth

The ground truth for evaluating the face clustering method was created by the authors, by annotating the 66,380 faces detected in the RCO dataset. The true P/NP labels were derived using synchronized symbolic information. As for the RCO dataset, we used four MIDI files synchronized to the video files provided by Grachten et al. [13], from which we extracted the P/NP labels with the method described in [5]. The Music Technology Group (Pompeu Fabra University, Spain) provided us with the video recording and a set of files encoding synchronized note onsets and offsets for each instrumental part. In both cases, each performing musician was bound to the corresponding instrumental part / MIDI track in order to build the corresponding ground truth P/NP sequence.

7.5. Evaluation approach

The goal of the experimental evaluation in this paper was threefold. First, we assessed the performance of the key-modules of our framework, including P/NP labeling (Section 8.2) as well as face labeling – i.e., musician diarization (Section 8.1). The quality of P/NP label sequences is the key result serving to demonstrate the effectiveness of our proposed method. However, we also evaluated the face labeling step to understand how inevitable errors there affect the quality of P/NP label sequences.

Second, as reported in Section 8, we assessed the quality of the obtained P/NP label sequences also relatively, using a random baseline as a reference. Relying on a random baseline was the only possible choice here, and this for the following reasons. The related literature does not offer a solution for yielding one sequence of P/NP labels for each performing musician. In fact, as discussed in Section 3, existing audio-based and visual-based classifiers cannot be directly applied to the type of audio-visual content considered in this paper. Replacing the semi-automatic framework modules described in Sections 6.3 and 6.4.2 is only theoretically possible. As explained in Section 3.1.4, existing vision-based classifiers require input of a particular type and are instrument-dependent.

Finally, in Section 8.6, we compared the efficiency of P/NP labeling using our method with the efficiency of the purely manual P/NP labeling in order to determine how much human annotation can be speeded up, while maintaining the same high quality of the P/NP label sequences.

7.6. Evaluation measures

In this section, we describe the evaluation measures used to assess the quality of the labels produced after the human annotation steps described in Sections 6.4.1 and 6.5.

Once the face clusters had been generated and labeled, we jointly evaluated precision, recall and number of labeled (or non-discarded) face detections. The average precision and the average recall were combined together into the average F1-score. The percentage of non-discarded face detections was simply determined by counting how many images inherit a label from non-discarded face clusters.

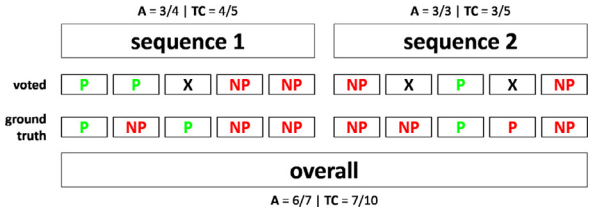


Fig. 13. Example of how A and TC scores are computed for P/NP label sequence assessment.

For each musician the system produces a sequence of P/NP/X labels to be compared to the corresponding ground truth sequence. As illustrated in Fig. 13, we evaluated the labeling performance integrally aggregating the results obtained for all the musicians. The performance with respect to the grond truth was assessed using two scores: *accuracy* (A) and *timeline coverage* (TC). The former is defined as the percentage of matching labels and it is computed only considering the known labels, namely those for which the value is different from X. The TC score is defined as the ratio between the number of non X-valued labels and the ground truth sequence length. It is an indicator of how many detections are used by the system and its upper bound is defined by the percentage of available musician detections.

We recommend to use accuracy instead of other scores, like precision and recall, because we need to assess how well the system produces both playing and non-playing labels. The timeline coverage was chosen to observe how many labels are effectively generated by the system, but also to measure the impact of rejecting non-pure image clusters.

8. Results

This section reports the results and provides the reader with the answers to the research questions defined in Section 1. First, we addressed RQ1 in Section 8.1, where we evaluated different options to solve the musician diarization problem. Then, in Section 8.2 we focused on the P/NP labeling problem addressing RQ2 and RQ3. We added a failure analysis section (Section 8.3) in which we explained how the system fails. This provides insights about the informativeness of static images (RQ4). The results obtained when adopting the two proposed strategies dealing with missing observations are reported in Section 8.4. Then, we qualitatively compared the ground truth and the generated P/NP sequences using the OSV dataset (Section 8.5). Finally, we answered RQ5 by measuring the achieved efficiency and effectiveness of the human annotation tasks (Section 8.6).

8.1. Face labeling

We evaluated the proposed semi-automatic method producing face labels on the RCO test set. This set consists of 64,805 detections belonging to 54 musicians. With these detections we generated 191,745 cannot-link constraints (see Section 6.2).

Fig. 14 shows that the most informative regions are the face and the upper body. Including scene information does not significantly improve the performance and the same holds for the cannot-link constraints. While including scene information did not impact the computation time, running the constrained version of k-means led to a significantly longer execution time. In general, we see that our method generates face labels with high average precision and recall. However, this result was not obtained just via the face clustering step but also using the human annotators' ability to discarding non-pure clusters. In fact, in the best case we already observe that about 20% of the detections fell into discarded face clusters. This means that a part of the clusters was not sufficiently pure.

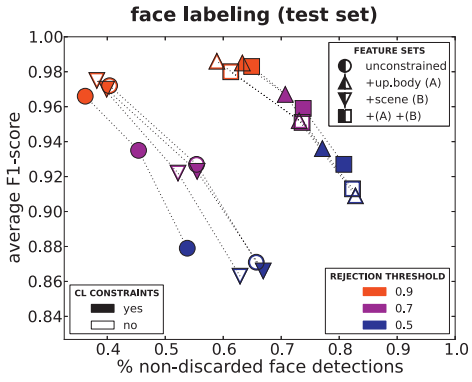


Fig. 14. We compared four feature sets (represented by different markers) using either the constrained clustering (filled markers) or the unconstrained one (empty markers). We also evaluated three different rejection thresholds (different colors). The plot shows three results. First, combining face and upper body visual information produces the best results. Second, adding scene visual information and/or using the cannot-link constraints does not significantly improve. Third, a higher rejection threshold effectively filters out non-pure clusters.

8.2. P/NP labeling

This section analyzes and compares the results obtained for the RCO and the OSV datasets. The research questions RQ2 and RQ3 were addressed in Section 8.2.3.

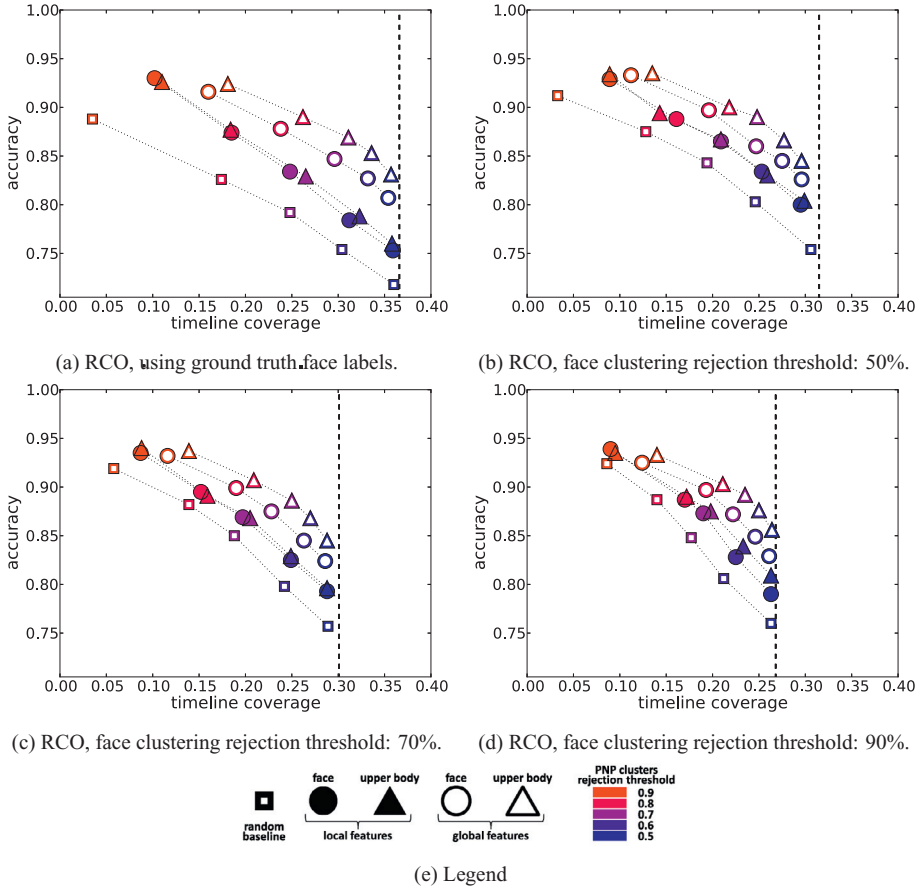
The plots reported in Sections 8.2.1 and 8.2.2 show the accuracy and the timeline coverage for the different types of features and regions of the image also including the results obtained with the random baseline method. As for the adopted notation, each point corresponds to the combination of an image region (upper body vs face), of type of features (global vs local) and of rejection threshold used when labeling the P/NP clusters (50, 60, 70, 80 and 90%). A dedicated marker is used for the random baseline method.

8.2.1. Evaluation on the RCO dataset

The RCO dataset allowed us to assess the full system that is, we could observe how different ways of generating the face labels affected the performance at the P/NP labeling step. To this end, we evaluated four cases. First, we considered the case of *ideal* input, in which the ground truth face labels were used. Then, we considered three different ways of obtaining the face labels by varying the rejection threshold used to label the generated face clusters. More specifically, we used the RCO test data, which includes the detections of 52 musicians. Setting α to 15 and approximating the entities set size $|E|$ to the number of musicians generated 780 face clusters. Then we simulated the annotation using three different rejection thresholds: 50% (tolerant annotator), 70% and 90% (strict annotator). In this experiment we used the unconstrained context-assisted face clustering method – i.e., we exploited face similarity and context information extracted from the upper body and the scene (see Section 6.2). The overall numbers of generated P/NP clusters were 530, 384, 354 and 342 for the face labels input of the types "ideal", 0.5, 0.7 and 0.9, respectively. In Fig. 15, which summarizes the results, we observe four facts.

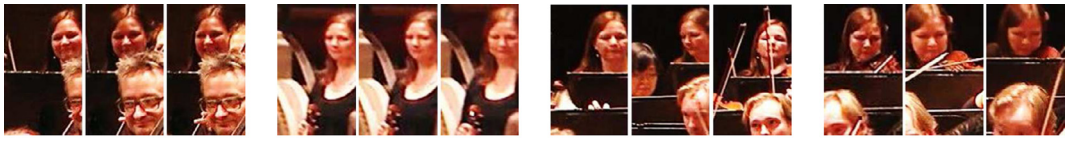
First, regardless of the input to the P/NP clustering step, there is a consistent trade-off between accuracy and timeline coverage. The stricter the annotator is (higher P/NP rejection threshold), the lower the number of produced P/NP labels is. More in detail, the figures show that the timeline coverage decrease is much larger than the accuracy increase. This means that quite often the purity of the produced P/NP clusters is below the highest rejection thresholds. In Section 8.3 we investigate the reasons why the P/NP clusters are not always pure enough.

Second, global features always outperform local ones and the upper body region is more informative than the face region. What is



(a) RCO, using ground truth face labels.
(b) RCO, face clustering rejection threshold: 50%.
(c) RCO, face clustering rejection threshold: 70%.
(d) RCO, face clustering rejection threshold: 90%.
(e) Legend

Fig. 15. Evaluation of the P/NP labels produced by the system. The vertical dashed lines show the upper bound for the timeline coverage, which is limited by the availability of face detections. The upper body region described with global features outperforms other combinations. Tuning the system for very high accuracy has a large negative impact on the timeline coverage. This shows that discriminating playing and non-playing HOIs requires information beyond a global description of a static upper body image.



(a) End of a movement (b) End of the concert (c) Non-playing (d) Playing

Fig. 16. Informativeness of the face region: even when the torso region is not visible we can guess whether a musician is playing by analyzing the face expression.

surprising is that faces are already a good indicator to infer P/NP labels. The advantage of this image region over the upper body one is that occlusions here seldom occurs. When the instrument or the human body parts are not visible, face cues can be always exploited. To show this, we give an example in Fig. 16. A relaxed, unfocused, or contemplative expression (Fig. 16a–c) is likely to be linked to a non-playing action, as opposed to a concentrated one (Fig. 16) that is likely to indicate a playing activity.

Third, when the rejection threshold for the sub-cluster annotation is set to 50%, the timeline coverage in Fig. 15 is always close to its upper boundary (the markers in the four plots are close to the vertical dashed lines). Such boundary is determined by the available face detections and it shows the highest possible timeline coverage. This result was expected because, by setting the rejection threshold to 50% and having only two possible labels (P and NP), no cluster is discarded. Still, a number of additional X labels can be generated by the process explained in Section 6.5 due to conflicting cluster labels in case of multiple views on the same musician. However, the plots reveal that this seldom happens.

Finally, by setting again the rejection threshold to 50%, we also observe that the accuracy is always above 75%. This happens because the numbers of P and NP labels in the ground truth are not equal. For this reason, in order to assess whether the proposed method is generating P/NP clusters at all, the random baseline method is included. What we see is that the baseline always performs worse, both in terms of timeline coverage and accuracy. This shows that our method effectively discriminates playing and non-playing actions.

8.2.2. Evaluation on the OSV dataset

In the OSV dataset, 63 musicians are recorded by a fixed camera. Compared to RCO, there is no point-of-view variability and all the musicians are always visible. The number of P/NP clusters is 126. For this recording we only evaluated local and global features extracted from the upper body region. The made this choice because, as explained in Section 7.3.2, the face region in the OSV dataset is too small. Even we could not use this recording to evaluate the full proposed system, it is an additional test case to also assess whether and to what extent other recordings and recordings of a different type can be exploited for P/NP detection.

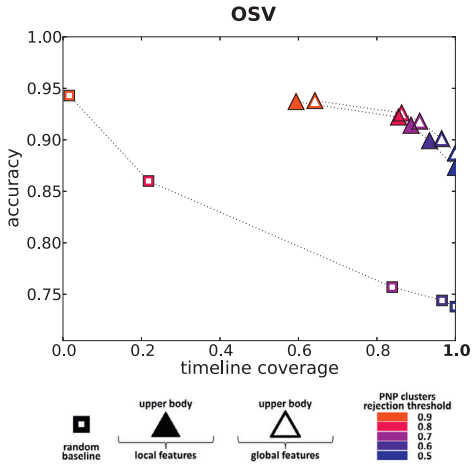


Fig. 17. Results for the OSV dataset. Even if the video resolution is low, with a fixed camera we accumulate a sufficient number of images for each musician from the same point of view. Due to this, playing and non-playing images can be discriminated with high accuracy.

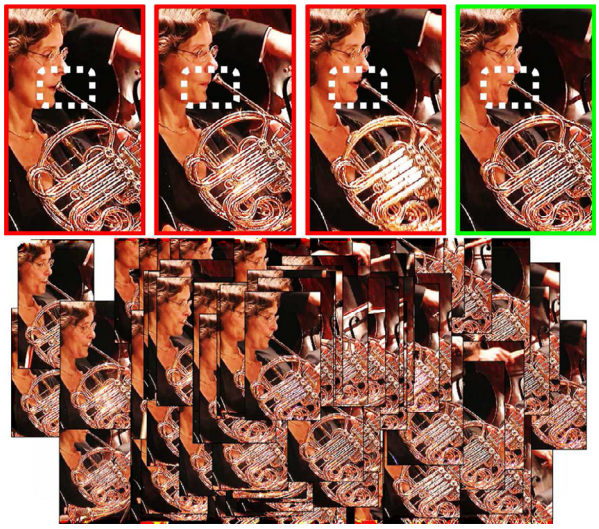


Fig. 18. Good P cluster containing some NP images, which are included because the differences in the mouth region are not dominant, sufficiently influencing the cluster formation.

The OSV images are challenging because they have a low resolution. However, the system is still able to well discriminate P/NP actions as shown by the results in Fig. 17. This becomes evident by comparing the random baseline performance with that of our image clustering methods. Increasing the rejection threshold from 50% up to 80%, we see that the number of discarded images decreases linearly at relatively small steps. This means that the majority of the generated P/NP clusters were pure enough. However, when we look at the strictest rejection threshold, we observe that the accuracy increase is small while the number of determined P/NP labels decreases at a much higher rate. Therefore, as we did for the RCO dataset, we conclude that there are additional factors determining the playing/non-playing status of musicians which are not taken into account in our solution.

8.2.3. Overall judgment

By evaluating on the RCO and the OSV datasets, we answered to the research questions RQ2 and RQ3.

We conclude that the most P/NP discriminative region is the upper body. However, we remark that faces by themselves are already surprisingly informative. Regarding the accuracy of the system, we see that it ranges between 70 and 94% depending on the strictness of the annotators. However, targeting to a high accuracy has a significant impact on the number of discarded detections especially in the case of a multiple-camera recording in which it is hard to continuously accumulate observations over time for each performing musician.

As for the impact of different modules, we have two conclusions. First, we see that the overall timeline coverage is directly affected by the number of available face detections. This indicates that the face detectors should be tuned to perform with high recall in order to determine as many P/NP labels as possible for each musician. Second, we observe that the musician diarization module has a limited impact on the overall accuracy because most of the face clusters are sufficiently pure.

8.3. Failure Analysis

As pointed out in Sections 8.2.1 and 8.2.2, a fraction of the produced sub-clusters S_m^c is not sufficiently pure. By inspecting the produced P/NP clusters, we found that subtle discriminative cues in the images sometimes occur. For instance, in Fig. 18, we see that the mouth region for the French horn player is the discriminative region. However, our method has not been designed to explicitly take into account this part of the image, therefore the images are clustered according to the overall appearance of the upper body.

The aforementioned error belongs to a larger class of errors, namely the false positives. By inspecting the videos, we observed that they occur for any type of instrument and that they are caused by *anticipation*, which occurs when a musician gets ready to play in advance. This is also supported by the confusion matrices in Fig. 19. They all show that the amount of false positives (false P labels) is greater than the amount of false negatives (false NP labels). Even if the P/NP ground truth has been generated taking into account anticipation [5], the results reported in Fig. 19 let us believe that it starts much earlier than expected.

Due to the aforementioned observations, we answer to RQ4 as follows. On the one hand, a more detailed analysis of the images can be performed (e.g., including features extracted by the mouth region) thanks to which a static image could be enough for P/NP labeling. On the other hand, we cannot exclude that an image itself is partially informative. For instance, we expect that musicians' movements could be informative as well. Additionally, timbral features from the audio recording can be used in a multimodal fashion.

8.4. Evaluating the strategies for missing detections

In Section 6.6 we proposed two ways of dealing with the limited availability of observations (namely, highest TC and merging). We evaluated the two strategies by considering the case of ideal face clustering input, using global features extracted from the upper body region and by setting the P/NP clustering rejection threshold to 80%. The results summarized in Table 1 show that both strategies are beneficial. In fact, when nothing is done (standard case), the timeline coverage is always the lowest.

In the highest TC case, the result is a direct consequence of using the labels from the most visible musician. While in the merging strategy, the advantage comes from the availability of multiple P/NP labels obtained by exploiting the musician redundancy within each instrumental part. Due to this redundancy, the voted labels can be inferred with more confidence at the majority voting step (see Section 6.5). Overall, the most effective strategy is merging.

8.5. Qualitative assessment

We also qualitatively assessed the P/NP labeling performance generating a PNP matrix. This matrix shows all the P/NP sequences

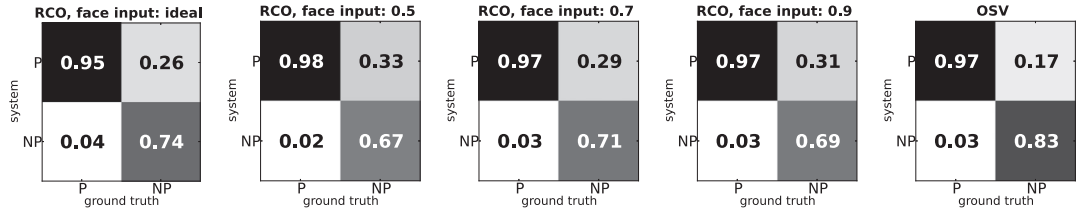
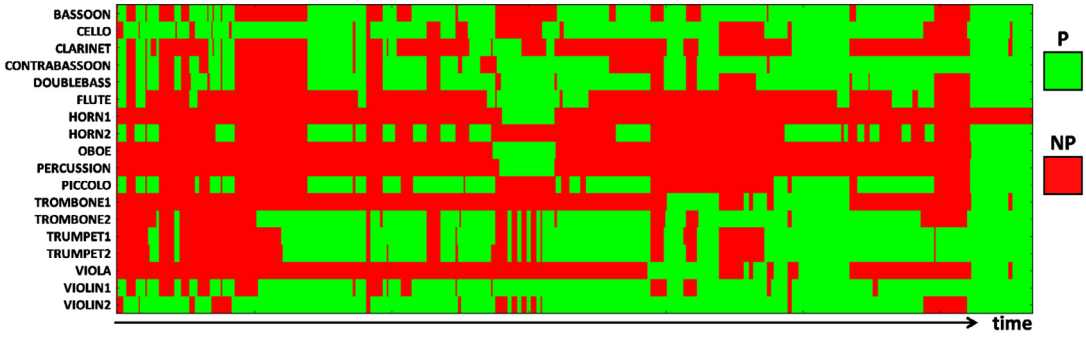
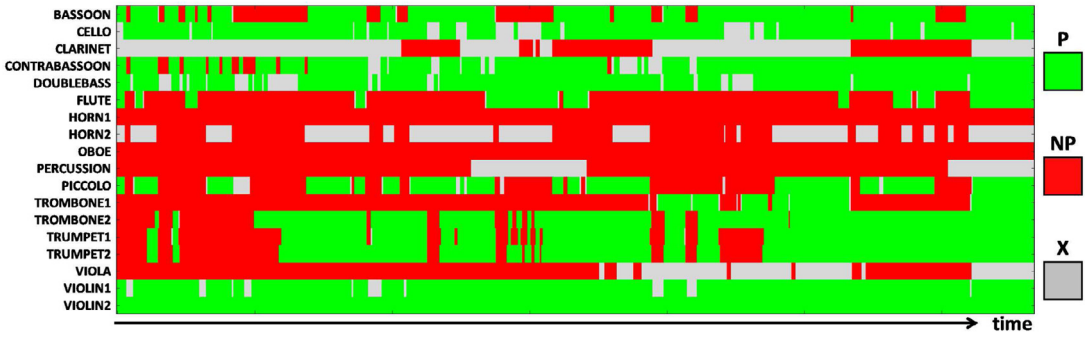


Fig. 19. The depicted confusion matrices show that the system has a bias towards false positives. Such a bias can be explained by the fact that the musicians usually get ready to play sufficiently in advance.



(a) Ground truth P/NP matrix.



(b) P/NP matrix via P/NP clustering

Fig. 20. Comparing the P/NP matrices for the OSV performance. The merging strategy has been applied. Therefore both matrices have one row per instrumental part.

Table 1
Comparing the standard method with two possible strategies dealing with missing observations. The scores are computed considering the ground truth face labels, global features extracted from the upper body region and adopting 80% as rejection threshold for the PNP clusters. The merging strategy significantly improves the performance in the RCO case, while it has limited benefit in the OSV case. This is expected since the latter is a fixed camera recording and every musician is always visible.

Strategy	Standard		Highest TC		Merging	
Score	A	TC	A	TC	A	TC
RCO	0.890	0.262	0.884	0.369	0.884	**0.429**
OSV	0.926	0.863	0.927	0.867	0.927	**0.873**

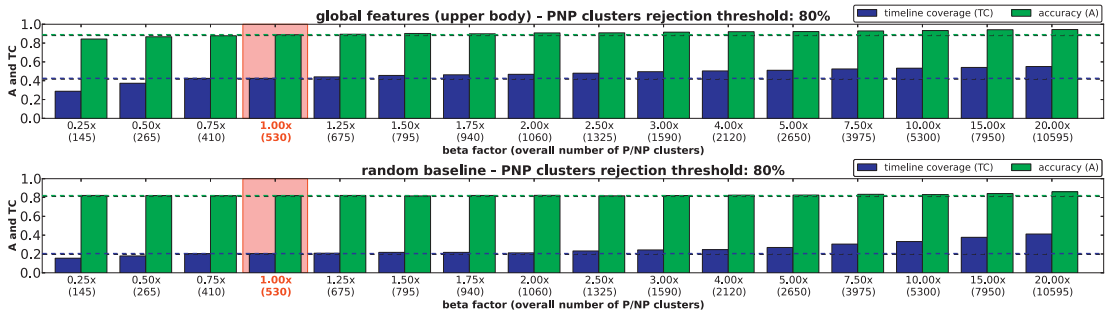
produced for different instrumental parts. In Fig. 20 compares the ground truth matrix and the one generated for the OSV dataset. The latter is generated using global features extracted from the upper body region and by setting the P/NP clustering rejection threshold to 80%.

From this example, we observe that the dominant error is indeed caused by the false positives and that for some instrumental parts a significant number of labels are missing (in particular for the clarinet and the horn).

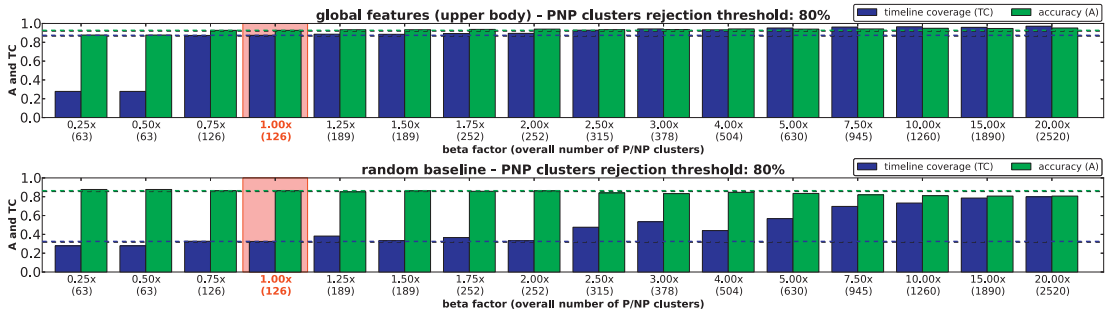
8.6. Human annotation efficiency

We addressed the last research question (RQ5) by assessing the balance between the efficiency and effectiveness of the human annotation required by our system. We evaluated whether the system generates a close-to-optimal number of P/NP clusters and measured the ratio between the amount of required human annotations and the number of generated P/NP labels. As for the notation used here, we refer to Section 6.3.

First, we assessed whether our method produces too many or too few sub-clusters S_m^c. Fig. 21a and b report the results for the RCO and the OSV dataset, respectively. The plots show how the timeline coverage (TC) and the accuracy (A) change by varying the number of generated P/NP clusters. The results were obtained using the P/NP clustering based on global features extracted from the upper body region. To show the significance of the results, we also evaluated the random baseline's performance. In both cases, the ground truth face labels are used and the P/NP clustering rejection threshold is set to 80%. For each musician we estimated the number of points of view and we considered twice as many sub-clusters (as explained in Section 6.3). Then, we used an additional factor β applied to increase (or decrease) the number of sub-clusters per musician. For instance, when $\beta = 5$, the number of sub-clusters is ten times the number of the estimated points of view. When $\beta = 0.5$ the number of sub-clusters is exactly



(a) RCO dataset (face ground truth labels used). By increasing the number of generated P/NP clusters ($\beta > 1$) both the A and the TC scores slightly increase (saturation of the performance). By contrast, when $\beta < 1$ our method generates much less pure P/NP sub-clusters. This is an indicator that a suitable number of sub-clusters is chosen. Differently, in the case of the random baseline method, increasing β leads to a substantial increase of the TC score. This shows that the number of generated sub-clusters is optimal for the method we propose.



(b) OSV dataset. In this case, there is only one point-of-view on every musician and therefore the estimated number of sub-clusters is 2 for every musician. When $\beta \in \{0.25, 0.50\}$, only one sub-cluster per musician is generated. Due to the P/NP rejection threshold set to 0.8, only those musicians who play for at least 80% of the performance timeline will be labeled as always playing. For this reason, we observe a drop of the TC score and a decrease of the A score. Differently, when $\beta > 1$, the performance slightly improves. This saturation shows once again that the dominant difference in the images is the performed playing/non-playing action and therefore that two sub-clusters for each point-of-view are enough.

Fig. 21. Assessing whether the amount of required human annotation by our system is optimal. We verify whether the system generates the optimal number of P/NP sub-clusters. Generating too many clusters leads to unnecessary human labor, on the other hand the critical number of P/NP sub-clusters has to be generated in order to avoid too many non-pure sub-clusters. We have added the horizontal dashed lines to compare the performance obtained by different values of β to that obtained when β is 1 – i.e., the default number of generated P/NP clusters.

the number of points of view. In summary, the value set for β affects the overall number of sub-clusters $\sum_{m=1}^{|M_{GT}|} C_m$.

In Fig. 21a, we see that on the left of $\beta = 1$, the performance quickly decreases. By contrast, on its right side the timeline coverage slowly increases. This pattern is even more evident for the OSV concert (Fig. 21b). In this case there is a sharp transition from the case in which there is only one sub-cluster per musician (namely when $\beta \in \{0.25, 0.50\}$) and a saturation of the performance for values of β bigger than the unity. Both results show that the way the system chooses the number of sub-clusters is optimal to avoid unnecessary over-segmentation. Adding too many clusters would lead to extra manual annotation but with little advantage in terms of accuracy and timeline coverage. Similarly, we see that the system generates the critical number of sub-clusters which are necessary to avoid that P and NP images consistently fall together into one cluster.

Finally, we computed the ratios between the overall number of detections and the number of produced sub-clusters. The former is defined as $\sum_{m=1}^{|M_{GT}|} |S_m|$, while the latter is defined as $\sum_{m=1}^{|M_{GT}|} C_m$. For the RCO dataset, the ratio is equal to $52204/530 = 98.5$ and for the OSV dataset $42084/126 = 334$. This means that on average one human label is propagated to about 100 detections in the RCO dataset and more than 300 in the OSV one.

9. Discussion

In this final section, we report the limitations we have encountered while deploying a number of state-of-the-art methods hence suggesting possible research directions for the future.

The face detection step is critical for our system since it directly affects the timeline coverage. We found that off-the-shelf detectors are optimized to achieve high precision and that the recall is not satisfying like evident, for instance, from the example of Fig. 6 in which approximately only one third of the musicians is detected. Our attempt to overcome this problem by combining multiple heterogeneous detectors helped, but it may be useful to investigate more how to improve the face detection recall in videos.

When clustering the faces, it is important to limit the number of generated clusters in order to reduce the amount of human annotation. State-of-the-art face clustering solutions designed to limit the number of produced clusters are available. However, they work assuming that the initial clusters are nearly 100% pure. What we found in our experiments is that this does not always hold. More specifically, we observed either very pure face clusters or fuzzy ones and that the latter usually contain images with lower resolution and/or profile faces. In order to maximize the utility of each detected face, and once again avoid negative impact on the timeline coverage, face

clustering methods should be improved so that non-pure clusters are detected and discarded or treated with alternative strategies.

Furthermore, a more detailed analysis of the image segmentation process is needed. The idea of exploiting the head pose to determine the upper body region of a musician seems to be effective. We evince this by inspecting the results obtained at the P/NP clustering step when upper body images are clustered – i.e., empirical evaluation of the segmentation process. However, it may be the case that the optimal size of the upper body bounding box changes for different types of instruments. Hence, a more detailed analysis of the segmentation process should be carried out, eventually measuring the performance in analytical fashion rather than an empirical one.

Finally, by investigating the limitations of our approach, we learned that there are cases in which a non-playing image is very similar to a playing one due to the anticipation before the actual note onsets. What we have observed shows that the playing/non-playing information is not simply encoded in the spatial configuration between the musical instruments and the body parts as assumed by state-of-the-art methods. Additional information has to be extracted by, for instance, exploiting the richness in the face region, the musicians' movements and/or auditory features. A second issue to be considered is how to possibly label the discarded images. For instance, using the non-discarded, and hence labeled, clusters of images, ad hoc classifiers could be trained to relabel the discarded face detections and the images from the discarded sub-clusters. Future work may also be directed towards the exploration of the additional information resources mentioned above and the exploration of relabeling strategies.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7 / 2007–2013 through the PHENICX project under Grant Agreement no. 601166. We would like to thank: the Royal Concertgebouw Orchestra and the Simfònica del Vallès Orchestra for providing us with their video recordings, Maarten Grachten (Austrian Research Institute for Artificial Intelligence, Austria) and the Music Technology Group (Pompeu Fabra University, Spain) for the synchronized scores, Giuseppe Lisanti (Media Integration and Communication Center, Italy) and Michael Riegler (Simula Research Laboratory, Norway) for their precious hints to realize this work.

References

[1] E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, A comparison of extrinsic clustering evaluation metrics based on formal constraints, Inf. Retr. 12 (4) (2009) 461–486.
[2] A. Arzt, G. Widmer, S. Dixon, Automatic page turning for musicians via real-time machine listening, in: Proceedings of the ECAI 2008-18th European Conference on Artificial Intelligence, Patras, Greece, July 21-25, 2008, 2008, pp. 241–245.
[3] A.D. Bagdanov, A.D. Bimbo, F. Dini, G. Lisanti, I. Masi, Posterity logging of face imagery for video surveillance, IEEE MultiMedia 19 (4) (2012) 48–59.
[4] J.G.A. Barbedo, G. Tzanetakis, Musical instrument classification using individual partials, IEEE Trans. Audio Speech Lang. Process. 19 (1) (2011) 111–122.
[5] A. Bazzica, C.C.S. Liem, A. Hanjalic, Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music, in: Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, 2014.
[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceeding of Workshop on statistical learning in computer vision, European Conference on Computer Vision (ECCV), 1, 2004, pp. 1–2.
[7] A. D'Aguanno, G. Vercellesi, Automatic music synchronization using partial score representation based on IEEE 1599, J. Multimed. 4 (1) (2009) 19–24.
[8] S. Dixon, W. Goebl, G. Widmer, The air worm: an interface for real-time manipulation of expressive music performance, in: Proceedings of the International Computer Music Conference, 2005, pp. 614–617.
[9] E. el Khoury, C. Sénac, P. Joly, Audiovisual diarization of people in video content, Multimed. Tools Appl. 68 (3) (2014) 747–775.
[10] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, 1996, pp. 226–231.
[11] S. Ewert, B. Pardo, M. Müller, M.D. Plumbley, Score-informed source separation for musical audio recordings: an overview, IEEE Signal Process. Mag. 31 (3) (2014) 116–124.
[12] E. Gómez, M. Grachten, A. Hanjalic, J. Janer, S. Jorda, C.F. Julia, C. Liem, A. Martorell, M. Schedl, G. Widmer, PHENICX: performances as highly enriched and interactive concert experiences, in: Proceedings of the Sound and Music Computing Conference (Stockholm, 2013), Citeseer, 2013.
[13] M. Grachten, M. Gasser, A. Arzt, G. Widmer, Automatic alignment of music performances with structural differences, in: Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013, 2013, pp. 607–612.
[14] C. Joder, S. Essid, G. Richard, A comparative study of tonal acoustic features for a symbolic level music-to-score alignment, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA, 2010, pp. 409–412.
[15] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2005.
[16] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17–22 June 2006, New York, NY, USA, 2006, pp. 2169–2178.
[17] M. Lux, S.A. Chatzichristofis, Lire: lucene image retrieval: an extensible java CBIR library, in: Proceedings of the 16th International Conference on Multimedia 2008, Vancouver, British Columbia, Canada, October 26-31, 2008, 2008, pp. 1085–1088.
[18] L.G. Martins, J.J. Burred, G. Tzanetakis, M. Lagrange, Polyphonic instrument recognition using spectral clustering, in: Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007, 2007, pp. 213–218.
[19] K. McGuinness, O. Gillet, N.E. O'Connor, G. Richard, Visual analysis for drum sequence transcription, in: Proceedings of the European Signal Processing Conference (EUSIPCO), 2007, pp. 312–316.
[20] M. Müller, S. Ewert, Joint structure analysis with applications to music annotation and synchronization, in: Proceedings of the 9th International Conference on Music Information Retrieval, ISMIR 2008. Drexel University, Philadelphia, PA, USA, September 14-18, 2008, 2008, pp. 389–394.
[21] M. Müller, N. Jiang, A scape plot representation for visualizing repetitive structures of music recordings, in: Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012, 2012, pp. 97–102.
[22] M. Müller, F. Kurth, T. Röder, Towards an efficient algorithm for automatic score-to-audio synchronization, in: Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, ISMIR 2004. Spain, October 10-14, 2004, 2004.
[23] M.H. Nguyen, Z. Lan, F.D. la Torre, Joint segmentation and classification of human actions in video, in: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011, 2011, pp. 3265–3272.
[24] C. Raphael, Aligning music audio with symbolic scores using a hybrid graphical model, Mach. Learn. 65 (2-3) (2006) 389–409.
[25] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: Proceedings of the IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009, 2009, pp. 1593–1600.
[26] D. Sculley, Web-scale k-means clustering, in: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, 2010, pp. 1177–1178.
[27] C. Shan, Face recognition and retrieval in video, in: Video Search and Mining, 2010, pp. 235–260.
[28] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, 2014, pp. 568–576.
[29] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, CoRR (2012). abs/1212.0402, online available http://crcv.ucf.edu/papers/UCF101_CRCV-TR-12-01.pdf
[30] T.F. Tavares, G. Odowichuck, S. Zehtabi, G. Tzanetakis, Audio-visual vibraphone transcription in real time, in: Proceedings of the 14th IEEE International Workshop on Multimedia Signal Processing, MMSP 2012, Banff, AB, Canada, September 17-19, 2012, 2012, pp. 215–220.
[31] V. Thomas, C. Fremerey, M. Müller, M. Clausen, Linking sheet music and audio - challenges and new approaches, in: Multimodal Music Processing, 2012, pp. 1–22.
[32] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[33] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, N.Q.K. Duong, The signal separation evaluation campaign (2007–2010): achievements and remaining challenges, Signal Process. 92 (8) (2012) 1928–1936.
[34] P.A. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA, 2001, pp. 511–518.
[35] C. Vondrick, D.J. Patterson, D. Ramanan, Efficiently scaling up crowdsourced video annotation - a set of best practices for high quality, economical video labeling, Int. J. Comput. Vision 101 (1) (2013) 184–204.

[36] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, Constrained k-means clustering with background knowledge, in: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, 2001, pp. 577–584.
[37] B. Wu, Y. Zhang, B. Hu, Q. Ji, Constrained clustering and its application to face clustering in videos, in: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, pp. 3507–3514.
[38] B. Yao, F. Li, Grouplet: A structured image representation for recognizing human and object interactions, in: Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010, 2010, pp. 9–16.
[39] B. Yao, J. Ma, L. Fei-Fei, Discovering object functionality, in: Proceedings of IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, 2013, pp. 2512–2519.
[40] L. Zhang, D.V. Kalashnikov, S. Mehrotra, A unified framework for context assisted face clustering, in: Proceedings of International Conference on Multimedia Retrieval, ICMR'13, Dallas, TX, USA, April 16-19, 2013, 2013, pp. 9–16.
[41] L. Zhang, D.V. Kalashnikov, S. Mehrotra, Context-assisted face clustering framework with human-in-the-loop, IJMIR 3 (2) (2014) 69–88.
[42] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, 2012, pp. 2879–2886.