# Using multi-dimensional correlation for matching and alignment of MoCap and Video signals

Michele Buccoli, Bruno Di Giorgi, Massimiliano Zanoni, Fabio Antonacci, Augusto Sarti
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci 32 - 20133 Milano, Italy
name.surname@polimi.it

***Abstract*—Motion analysis and tracking often relies on multi-modal signals, e.g., video, depth map, motion capture (MoCap), due to the completeness of information they jointly provide. The joint analysis of multimodal signals requires to know the correct timing, i.e., the signals to be aligned. In this paper we propose an approach to automatically estimate the correct matching and alignment between a video and a MoCap recording acquired from the same session, based on the multi-dimensional correlation of velocity-based features extracted from the two recordings. We validate our approach over a dataset of dance recordings of four genres, and we achieve promising results for both the alignment and matching scenarios.**

## I. Introduction

Motion tracking and motion analysis have been recently receiving a great deal of attention, due to their numerous application scenarios [1], spanning from surveillance to games, medicine for orthopedic patients or monitoring of athletes [2]. The art of dance might highly benefit from motion analysis as well [3], e.g., for the automatic annotation of dance performances, the definition of similarity metrics among steps, which would lead to the ultimate goal of assisting dance teachers and students and provide powerful tools for choreographers [4].

Many approaches for motion analysis are based on video signals, which are easy to capture and whose analysis has proved to achieve good results on the task [5][6][7]. However, the video signals lack information on depth and therefore the applications based on videos have a limited extent [8][9]. This is the reason that leads many studies to analyze RGB-D videos, which also provide the third dimension for depth, and Motion Capture (*MoCap*) recordings. The latter, in particular, provide a high amount of information, given by the 3D positions and rotations of numerous body joints [8][9], displayed as a 3D skeleton. While some techniques to extract skeleton information directly from RGB-D data have been proposed [8][9], the *ad-hoc* solutions based on optical markers, such as Vicon or Qualisys systems, are the state of the art for the task.

In several applications the analysis on different types of recordings (audio, video, MoCap) of the same performance (multimodal analysis) allows to achieve better result due to the variety of information that it is possible to collect from the signals [2][10][11][12]. In order to perform such multimodal analysis, it is crucial the signals to be *synchronized* [11][13][14][15]. Typically, the alignment of the signals is performed either before or after the data acquisition. In the case the alignment is performed before the acquisition, a common clock signal is delivered across the involved acquisition devices in order to have a common time reference [12]. In the case the alignment is performed after the acquisition, the recordings are manually annotated with the time offset, based on some kind of visual or audio cues.

Unfortunately, it is not always possible to take advantage of a clock signal. This is the case when some of the devices do not support an external clock. On the other side, the high number of recordings makes the manual annotation a cognitively-heavy task, and even the correct matching of the signals (i.e., which ones were captured from the same session) might not be available, e.g., due to unstructured recordings.

It is therefore necessary to develop an automatic technique to perform the matching among multimodal recordings of the same session and to estimate the correct offset for time alignment. In this paper we propose a method for matching and alignment of video and MoCap streams.

Automatic alignment has been successfully adopted for video/audio and video to video context [16], while audio matching (identification) is a common task in the Music Information Retrieval research field [17]. To the best of our knowledge no automatic method for video/MoCap matching and alignment has been proposed.

The method we propose is based on the correlation of features extracted from video and MoCap signals, which results in a metric of likelihood of the two signals to belong to the same session (matching) at a certain time offset (alignment). This task involves several steps, which are described in the first Sections of the paper. First, we need to identify the most feasible descriptors for the representation of the signals and to be compared across different multimodal signals (Sec. II). We use velocity-based features toward the horizontal and vertical directions, since they are comparable between MoCap and videos [10][18]. Then, the correlation needs to address many issues, such as noisy or poorly informative signals (Sec. III). Finally, we use the proposed approach for two use-case scenarios (Sec. IV). In the former scenario, we assume the matching between video and MoCap recordings is known and we perform the sole alignment between them. In the latter and more challenging scenario, we use the proposed method to perform both matching and alignment of the two sets of video and MoCap streams.

978-1-5090-3649-3/17/$31.00 2017 IEEE

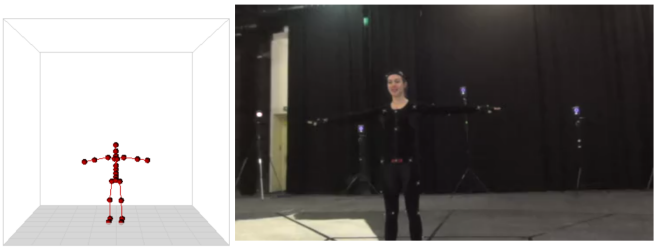

Fig. 1. An example of a video and MoCap recordings of the same session. Please note the slight misalignment between the camera's perspectives.

## II. Feature representation

The proposed approach requires to compare a video and a MoCap signals to estimate the likelihood they refer to the same session and their relative time offset. Video and MoCap signals are not directly comparable since they capture different information (i.e., light and color intensity for the video, positions in the MoCap), in different format. For this reason we adopt a feature-base representation.

We aim to extract a set of features that satisfy the following requirements:

- they need to be **consistent**, i.e., to represent the same physical quantity across the different signals, such as the position, velocity or acceleration of the actors in the scene;
- they should capture **global** properties for each frame, in order to be robust to the noise that might affect local properties of the recordings (such as body joints/limbs or single pixels);
- they must be **invariant** to a set of transformations, such as spatial alignment of the actors; for this reason, the position is not likely to be a reliable descriptor.

While the proposed algorithm might work for a variety of feature representation that satisfy the aforementioned requirements, in this study we adopt the physical velocity (i.e., the first-order derivative of the position) of the actor in the scene, which is independent by the position of the actor.

In the following, we describe the procedure to extract the features from the video (through the Optical Flow descriptor) and MoCap (through the first-order derivative of the position of the joints) signals.

### A. Video's Optical Flow

Given a video signal with $F$ frames and three color channels (RGB components), we first process it to remove the background and highlight the foreground, by performing a Gaussian blur filtering, with an isotropic Gaussian kernel of standard deviation $\sigma_b$, and a temporal FIR filtering, where the filter is the composition of a first order difference and a length $\eta_{ma}$ moving average. The filtering operation is performed separately for each RGB channel, in order to exploit all available information, and the results are merged as a weighted sum ($0.299R + 0.587G + 0.114B$). The first-order difference operation can return negative values, therefore we finally extract the magnitude of the movement by computing the absolute value of the signal.

We then extract the Dense Optical Flow, which detects the direction of the apparent motion of subjects present in the scene [19]. We use the Gunnar Farnebacks algorithm [20] to estimate the local apparent motion of each pixel by comparing its neighborhoods over pairs of successive frames using a quadratic polynomial approximation of them. If the background is static enough, so that it has been removed in previous processing stages, and the camera was steady during the recording, the Dense Optical Flow provides a reliable descriptor of the direction and the amount of movement in the scene, i.e., its velocity. We aggregate the velocities over pixels by computing the sum over vectors for each frame and we consider the two components of the vectors, which represent the global amount of movement toward the horizontal and vertical directions. We apply a final $\eta_{mm}$-long moving median filter to the velocity signals, in order to remove the spikes that are due to sudden scene or light changes. The procedure results in two velocity sequences, one for the horizontal and one for the vertical directions, collected in the matrix $\mathbf{X}^v \in \mathbb{R}^{2\times F-1}$ for a generic video $v$.

### B. MoCap Velocities

We formalize a generic MoCap signal composed of $F$ frames, $J$ markers (or body joints) and the 3 dimensions in space as the tensor $\mathbf{M} \in \mathbb{R}^{F\times J\times 3}$.

Since we want to build a velocity representation consistent with the one extracted from the video, we take into consideration the set of markers correspondent to the parts of the body that are clearly captured by the video, i.e., the head, the torso, the legs and the arms. We extract first-order derivatives for each marker and sum them over markers in order to obtain a global velocity descriptor $\mathbf{v}_f$ for each frame $f$:

$$\mathbf{v}_f = \sum_{j=1}^{J} (\mathbf{M}_{f+1,j,:} - \mathbf{M}_{f,j,:}) \qquad f = 1, \ldots, F-1. \quad (1)$$

We project the 3D velocity signal onto the projection plane of the camera [21] such that the two dimensions on the plane match the horizontal and vertical directions of the video features. Assuming that the position and orientation of the video camera (with respect to the 3D MoCap space) are known, the projection plane is defined by the span of the two unit vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. We project the velocity features $\mathbf{v}$ onto $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ to obtain the horizontal and vertical components, respectively, which are collected in the final matrix $\mathbf{X}^m \in \mathbb{R}^{2\times F-1}$ for a generic MoCap $m$.

It is worth remarking that we are not considering any clipping window [21], so the MoCap features are computed even when the video signal is occluded (e.g., the actor is outside the view of the camera and therefore no motion is detected). For this reason, our method includes a stage to detect the segments where a meaningful amount of information (i.e., movement) occur, as discussed in Section III-A.

### C. Oversampling of the signals to match the sample rate

The proposed approach is based on a multi-dimensional correlation among the features extracted from the two signals, which therefore need be re-sampled to the same frame-rate. Undersampling the higher-rate signal would lead to lose information and possibly introduce aliasing issues. For this reason, we rather oversample the lower-rate by using a linear interpolation. The systems for the MoCap recording commonly use a frame rate of around 60-250 Hz, that is two to four times higher than the time resolution of traditional video cameras (25-30 Hz), so the latter is the most likely candidate for the oversampling.

## III. Multi-dimensional correlation

Given a video and a MoCap stream and the corresponding velocity-based features, the proposed method estimates a (sorted) set of likelihoods that the two streams were acquired from the same session, and their relative offset. This is accomplished with a series of processing steps:

1) **segmentation**: we extract the segments that contain a meaningful amount of information (i.e., movement), in order to focus on them and possibly fasten the computation;
2) **multi-dimensional correlation**: we correlate all the pairs of video and MoCap segments on both horizontal and vertical dimensions, resulting in a list of estimates of offset and their corresponding reliabilities
3) **clustering**: we cluster the computed offset and update their values with a weighted mean based on reliability estimates.

### A. Segmentation

We want to pass to the next computation stage only the segments of the recording that contain a meaningful amount of movement. This is required since in the subsequent multi-dimensional correlation step, the velocity sequences are being normalized and therefore small movements could have a detrimental effect on the estimation. Both video and MoCap features follow the same segmentation procedure.

Given a generic matrix of features $\mathbf{X}$ with $F$ frames, we extract the magnitude of movement by aggregating the two horizontal and vertical components as indicated by $d$:

$$\mathbf{m}_f = \sqrt{\sum_d \mathbf{X}(d,f)^2} \qquad f = 1, ..., F-1 \tag{2}$$

and we threshold $\mathbf{m}$ with its $\pi$-th percentile $P_\pi$ obtaining the binary sequence $\tilde{\mathbf{m}}$, which is equal to 1 when $\mathbf{m} > P_\pi$ and 0 otherwise.

The result at this stage is over-segmented, therefore we apply the binary closing and opening morphology operators, in order to remove gaps smaller than $\tau_1$ and segments smaller than $\tau_2$. The segments containing meaningful movement are passed to the subsequent multi-dimensional correlation stage.

*using $i$ as estimate index and $k$ as cluster index*;
**foreach** *pair* $(\hat{l}_i,\ \hat{c}_i)$ **do**
  **if** $\min_k |\hat{l}_i - l_k| < l_{thr}$ **then**
    $k^* \leftarrow \arg\min_k |\hat{l}_i - l_k|$;
    $l_{k^*} \leftarrow \frac{l_{k^*} c_{k^*} + \hat{l}_i \hat{c}_i}{c_{k^*} + \hat{c}_i}$;
    $c_{k^*} \leftarrow (c_{k^*} + \hat{c}_i)$;
  **else**
    create a new cluster;
  **end**
**end**
**return** $\{(l_k,\ c_k)\}$ *sorted by* $c_k$ *(reverse order)*;

**Algorithm 1:** The clustering of offsets and reliabilities.

### B. Multi-dimensional correlation

Let $\mathbf{X}_p \in \mathbb{R}^{2\times F_p}$ be the sequence of features corresponding to the generic $p$-th segment.For each unique pair of segments $(p, q)$ from a MoCap ($p$) and video ($q$) stream, with $F_p$ and $F_q$ frames, respectively, we compute the multi-dimensional correlation as

$$c_{p,q}(l) = \sum_d \sum_{t\in\mathcal{T}} \frac{\mathbf{X}_p(d,t)\mathbf{X}_q(d,t+l)}{\boldsymbol{\sigma}_p(d)\boldsymbol{\sigma}_q(d)} \tag{3}$$

where $l$ is the lag index, representing the delay (in frames) of the MoCap stream with respect to the video stream; $\boldsymbol{\sigma}_p$ and $\boldsymbol{\sigma}_q$ are the standard deviations of the two feature sequences; and $\mathcal{T} = [\max(0, -l), \min(F_p, F_q - l)]$ is the set of time frames where the two segments overlap, given the offset $l$ (this is equivalent to zero padding the sequences to compute "full" correlation).

The offset that maximizes the correlation between the pair of sequences is then adjusted by considering the frame indexes $s(p)$ and $s(q)$, when the two segments $p$ and $q$ start:

$$\hat{l}_{p,q} = \arg\max_l c_{p,q}(l) + s(p) - s(q), \tag{4}$$

and we use

$$\hat{c}_{p,q} = \max_l c_{p,q}(l) \tag{5}$$

as an indicator of the reliability of the estimated offset.

In this step we estimate, for each pair of segments from the MoCap and video streams, the lag that maximizes the multi-dimensional correlation. In order to retrieve the most likely offset between the two streams, we need to cluster together the information inferred from the different pairs of segment.

### C. Clustering

In this last stage we cluster together the offsets having close values and compute the corresponding aggregated reliabilities (i.e., the correlation values as computed in Eq. 5) . Each cluster represents a set of offsets, which we use to compute the overall offset as the (weighted) centroid of the cluster. We compute the reliability of a cluster as the sum of the included reliabilities. The procedure is detailed in Algorithm 1.

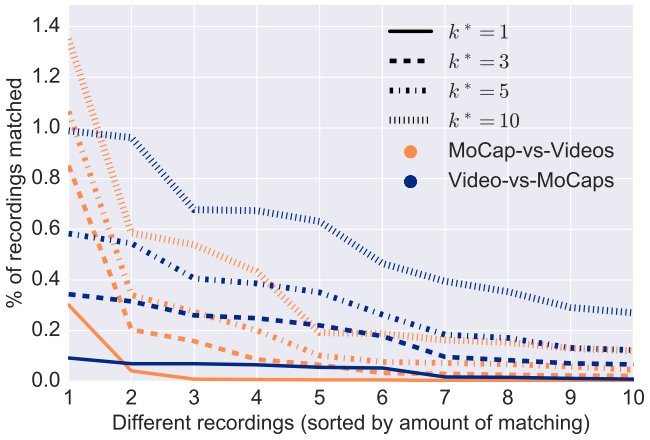


Fig. 2. Percentage of recordings matched at different estimates $k^*$ considered (style and width of the lines), for the 10 most matching recordings, using the MoCap-vs-Videos and Video-vs-MoCaps strategies (colors of the line). Best seen in color.

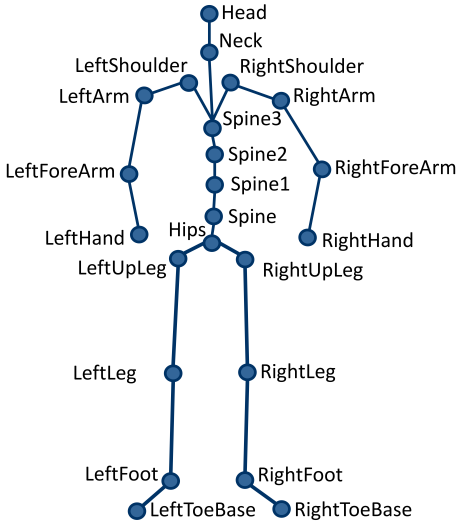


Fig. 3. Visualization of the joints in the MoCap representation

## IV. Alignment and Matching using the Multi-dimensional Correlation

In this Section we discuss how we use the proposed approach to estimate the correct alignment between two multimodal recordings of the same session, or to retrieve the best-matching MoCap recording given a video recording.

### A. Alignment and Matching

For the scenario of sole alignment, we assume to have a set of matched pairs of MoCap ($m$) and video ($v$) recordings of the same session and we need to automatically estimate the offset among those. Given a pair of recordings ($m$, $v$), we compute the set of reliabilities and offsets $(c^k_{m,v}, l^k_{m,v})$ sorted by decreasing reliability, where $k = 1, ..., K_{m,v}$ and $K_{m,v}$ is the number of estimates. The best candidate is $l^1_{m,v}$. In some scenarios an user might be interested in considering the remaining estimates $k > 1$, e.g., to analyze possible periodicities in the velocity signals.

For the scenario of matching, we define two alternative strategies, the MoCap-vs-Videos or the Video-vs-MoCaps. We here only discuss the former, while the latter is deducible from the duality of the approach. Given a generic MoCap stream $m$, we compute, for each video in the dataset, the set of estimates and we sort them across all the videos. The result is a set of pairs $(c^{k^*}_{m,v}, l^{k^*}_{m,v})$, with $k^*$ the new index given by the sorting across all the videos and $v$ a generic video. In this scenario, the reliability $c^{k^*}_{m,v}$ indicates the likelihood that the MoCap and video recordings have been acquired from the same session with the time offset $l^{k^*}_{m,v}$. The best video and offsets candidate are those corresponding $k^* = 1$, and the remaining estimates $k^* > 1$ might be used to retrieve similar recordings.

The matching scenario is extremely challenging, since we use the reliabilities as a global indicator of matching throughout the entire set of MoCaps or videos. However, the value of the reliability depends on many factors, such as the number of segments, the individual reliabilities of which it is composed, the duration of the involved recordings, the quantity of motion, etc. In the next Subsection we perform a preliminary test to analyze the effectiveness of the reliability to estimate the likelihood of matching and alignment.

### B. The hub issue

With regard to the MoCap-vs-Videos strategy, we collect the most popular videos, i.e., the videos that were matched with the highest amount of MoCaps, at different indices $k^*$. We perform the same analysis for the Video-vs-MoCaps strategy. The results are shown in Figure 2, where the amount of matched recordings is shown as a percentage of the total amount of recordings. With $k^* = 1$, the most popular recording is matched with around 10 to 30 % of the entire dataset. Due to the extreme popularity of such recordings, we name them *hubs*. At $k^* = 10$, the issue greatly affects the dataset, since the amount of recordings matched with the first hub is greater or equal than the size of the dataset itself. While this might seem an issue caused by a few outliers, we discover that even by removing them, other recordings are likely to take the lead role as hubs.

We address this issue by including a variant of the two matching strategies. With regard to the MoCap-vs-Videos strategy, we collect the whole set of $N \times K$ estimates across all the $N$ MoCap recordings, with $K$ the maximum amount of considered estimates across all the MoCap and video sequences. Then, we pick the estimate with the highest reliability $c^*_{m,v}$, we assign the MoCap recording $m$ to the video recording $v$ with the corresponding offset, and we remove from the set all the estimates regarding the MoCap $m$ and the video $v$. We iterate this procedure until the set of estimates is empty. Doing this, the video hubs are only matched with the most likely MoCap, and their impact on the whole matching procedure is dramatically reduced.

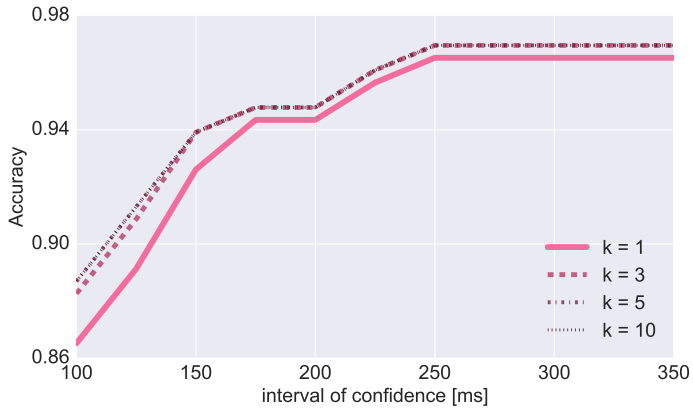


Fig. 4. Numerical evaluation of the alignment stage at different intervals of confidence and amount of estimates (style of the lines).

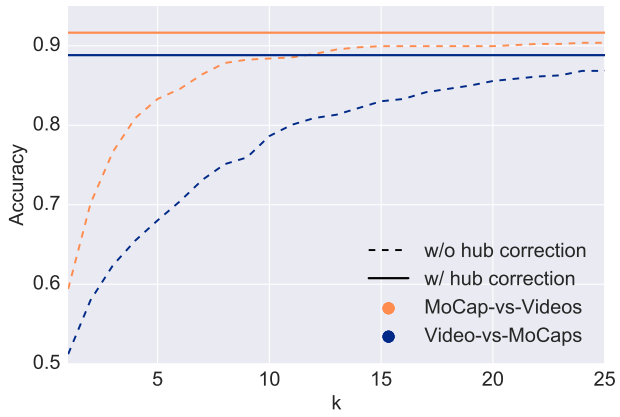


Fig. 5. Numerical evaluation of the matching stage at different amount of estimates, for the Mocap-vs-Videos and Video-vs-Mocaps strategies (color of the lines), with and without the strategy for hubs correction (style of the lines).

## V. Experimental Setup and Evaluation

### A. Collection of the dataset

We recorded $N = 707$ dance sessions from four different dance genres: Classic Ballet (130 sessions), Contemporary (347), Flamenco (49) and Green Folk dance (181).

For each session, we captured the performance of the dancers by means of a Vicon Motion Capture system[1], from which we built a skeleton-based representation composed of $J = 22$ joints (see Figure 3) at 60 fps. We recorded a video of the performance at 29 fps by means of a full-HD camera, approximately placed in front of the recording stage. In order to fasten the computation of the video features and focus to the movement of the dancers, we cropped the sides of the video and downscaled it. The final resolution was $480 \times 270$ pixels.

From the skeleton and the video recordings we computed the velocity-related features as explained in Section II. In particular, we computed $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ according to the position of the camera, while some minor changes of positions occur across the recordings (see for example Figure 1). Nevertheless, the results show that the features are rather robust to such minor misalignment.

We extracted the velocity features from the recordings, using the following setup as experimentally determined. The parameters for the initial video smoothing were set to $\sigma_b = 3$ and $\eta_{ma} = 4$. We uses the opencv[22] library to compute the optical flow features, and we set the parameters for the segmentation to $P_\pi = 35$, $\tau_1 = 0.1$ s and $\tau_2 = 0.23$ s.

### B. Collection of the ground truth

The manual annotation of the alignment between a video and a MoCap signal is a hard task. For this reason, we only considered the recordings of *Flamenco* and *Green Folk* sessions.We annotated the correct matching between MoCap and video recordings and their time offset with a precision of about 0.1 s.

[1]https://www.vicon.com

With regard to the *Contemporary* and *Ballet* sessions, we reverse the problem and we use the proposed approach to compute a set of estimates of matching and alignment. We then manually annotated the correct matching and offset candidates, if any.

### C. Evaluation of the sole alignment scenario

In Figure 4 we show the evaluation of the proposed approach for the task of the sole alignment, using the accuracy quality metric. We evaluate the accuracy at different intervals of confidence, i.e., the time interval from the annotated offset within which a value of alignment is marked as correctly retrieved. We also take into consideration the number of estimates required to find the correct offset. It is possible to note that even with $k = 1$, the achieved results are higher than 0.9 within 150 ms of interval of confidence. The additional estimates $k > 1$ slightly improve the performance, and the best accuracy is achieved within an interval of confidence of 250 ms.

We believe that the residual error is caused by the spatial misalignment among the features, or by those recordings that exhibit a small amount of movement. Nevertheless, the results are highly promising, which motivates the use of the proposed method to address the matching scenario.

### D. Evaluation of the matching scenario

In Figure 5 we show the accuracy achieved by the proposed approach for the two strategies at different $k$. It is clear that the MoCap recordings provide a higher discriminability with respect to the video recordings. In particular, using $k = 1$, the achieved accuracy is higher than 0.6, which improves to around 0.85 when considering the top 5 estimates. The Video-vs-MoCaps strategy achieves worse results, with only 0.5 accuracy for $k = 1$. It is clear that the hub issue described in Section IV-B negatively affects the overall accuracy of the approach.

This is highlighted by the results achieved by the two strategies when the variant for the hub correction is applied.

The results greatly improve and we achieve around 0.9 of accuracy for both the strategies. As before, the Mocap-vs-Videos strategy seems to slightly outperform the Video-vs-Mocaps one. The results are consistent with those obtained with the alignment scenario.

## VI. Conclusion

In this paper we addressed the problem of multimodal signal matching and alignment. In the specific scenario of dance performances we take into consideration video and MoCap acquisitions to be matched (find those that are related to the same session) and aligned. In particular, our method estimates the likelihood of MoCap and Video recordings to have been acquired from the same session, with a certain offset.

We validate our approach in two scenarios: alignment of pairs of streams (video-MoCap) already matched (we know they belong to the same session), but not aligned; matching and alignment of video and MoCap streams.

The approach, based on a multi-dimensional correlation of 2D velocity-based features achieves high accuracy (about $90\%$) for even narrow intervals of confidence for the alignment scenario. For the matching scenario, we design a procedure to use the approach over the global dataset, which results to be extremely effective. The approach can be used to assist researchers and artists in the organization of dataset of multimodal recordings.

As future works, we intend to remove the assumption of prior knowledge of the camera position, adding a stage for the estimation of the vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. This will widen the use-case scenarios of our approach to perform the matching of multiple video recordings of the same MoCap from different points of view. We will also explore further approaches to include a semantic layer of abstraction, such as multimodal deep learning techniques, which however will require a greater amount of data for training.

## Acknowledgment

This research activity has been funded by the European Project WhoLoDancE[2]. The authors would like to thank Katerina El Raheb and the other researchers and staff from Athena University for the help they provided during the collection of the ground truth.

## References

[1] T. B. Moeslund, A. Hilton, and V. Krger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 23, pp. 90 – 126, 2006, special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.

[2] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231 – 268, 2001.

[3] G. Volpe, A. Camurri, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, "The dancer in the eye: Towards a multi-layered computational framework of qualities in movement," in *Proc. of the 3rd International Symposium on Movement and Computing (MOCO)*, 2016.

[4] A. Camurri, K. El Raheb, O. Even-Zohar, Y. Ioannidis, A. Markatzi, J.-M. Matos, E. Morley-Fletcher, P. Palacio, M. Romero, A. Sarti, S. Di Pietro, V. Viro, and S. Whatley, "Wholodance: Towards a methodology for selecting motion capture data across different dance learning practice," in *Proc. of the 3rd International Symposium on Movement and Computing (MOCO)*, 2016.

[5] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1993–2008, Nov 2013.

[6] L. Sigal and M. J. Black, "Guest editorial: State of the art in image- and video-based human pose and motion estimation," *International Journal of Computer Vision*, vol. 87, no. 1, p. 1, 2009.

[7] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Proc. of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2013.

[8] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[9] L. Lo Presti and M. La Cascia, "3d skeleton-based human action classification," *Pattern Recognition*, vol. 53, no. C, pp. 130–147, May 2016.

[10] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe, *Multimodal Analysis of Expressive Gesture in Music and Dance Performances*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 20–39.

[11] M. Gowing, P. Kell, N. E. O'Connor, C. Concolato, S. Essid, J. Lefeuvre, R. Tournemenne, E. Izquierdo, V. Kitanovski, X. Lin, and Q. Zhang, "Enhanced visualisation of dance performance from automatically synchronised multimodal recordings," in *Proc. of the 19th ACM International Conference on Multimedia*, 2011.

[12] S. Piana, P. Coletta, S. Ghisio, R. Niewiadomski, M. Mancini, R. Sagoleo, G. Volpe, and A. Camurri, "Towards a multimodal repository of expressive movement qualities in dance," in *Proc. of the 3rd International Symposium on Movement and Computing (MOCO)*, 2016.

[13] M. Kulbacki, J. Segen, and J. P. Nowacki, *4GAIT: Synchronized MoCap, Video, GRF and EMG Datasets: Acquisition, Management and Applications*. Cham: Springer International Publishing, 2014, pp. 555–564.

[14] N. P. van der Aa, X. Luo, G. J. Giezeman, R. T. Tan, and R. C. Veltkamp, "Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," in *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011.

[15] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated humanmotion," *International Journal of Computer Vision*, vol. 87, no. 1, p. 4, 2009.

[16] S. Lameri, P. Bestagini, and S. Tubaro, "Video alignment for phylogenetic analysis," in *Proc. of the 24th European Signal Processing Conference (EUSIPCO)*, 2016.

[17] M. Zanoni, S. Lusardi, P. Bestagini, A. Canclini, A. Sarti, and S. Tubaro, "Efficient music identification approach based on local spectrogram image descriptors," in *Audio Engineering Society Convention 142*, 2017.

[18] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz, "Hmm-based human motion recognition with optical flow data," in *Proc. of the 9th IEEE-RAS International Conference on Humanoid Robots*, 2009.

[19] A. M. Tekalp, *Digital Video Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2015.

[20] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," *Image analysis*, pp. 363–370, 2003.

[21] E. Angel and D. Shreiner, *Interactive Computer Graphics: A Top-Down Approach with Shader-Based OpenGL*, 6th ed. USA: Addison-Wesley Publishing Company, 2011.

[22] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision in C++ with the OpenCV Library*, 2nd ed. O'Reilly Media, Inc., 2013.

[2] http://www.wholodance.eu