AUTOMATED BOBBING AND PHASE ANALYSIS TO MEASURE WALKING ENTRAINMENT TO MUSIC

A. López-Méndez [1], C.E.I. Westling [2], R. Emonet [3], M. Easteal [4], L. Lavia [5], H.J. Witchel [6], J-M. Odobez [7]

[1]: feezoo, Spain [2]: MFM, University of Sussex, UK [3]: Lab. Hubert Curien, France
[4]: Brighton & Hove Council, UK [5]: Noise Abatement Society, UK
[6]: BSMS, University of Sussex, UK [7]: Idiap Research Institute, Switzerland

ABSTRACT

In this paper, we investigate the influence of music on human walking behaviors in a public setting monitored by surveillance cameras. To this end, we propose a novel algorithm to characterize the frequency and phase of the walk. It relies on a human-by-detection tracking framework, along with a robust fitting of the human head bobbing motion. Preliminary experiments conducted on more than 100 tracks show that an accuracy greater than 85% for foot strike estimation can be achieved, suggesting that large scale analysis is at reach for finer music/walking behavior relationship studies.

***Index Terms*—** tracking, gait, bobbing estimation, entrainment to music

1. INTRODUCTION

In recent years there has been a growing interest in the video analysis community to go beyond the typical surveillance scenarios (e.g., detection of violence or left luggage) and to design algorithms to analyze a wider range of behaviors, such as group identification, dominance [1], or characterization of interpersonal relations (e.g. deception [2]). While such trends allow sociologist to envision the study of human behavior "in the wild" and thus in more ecologically valid settings, this also creates challenges, since the quantities of interest for performing such analyses are usually hard to estimate. One can cite for instance body and head orientation [3, 4] for social attention modeling, or the visual detection of voice activities [5].

Goal and motivation. We are interested in the influence of music on human walking behaviors and propose video processing as a method for automatically characterizing such behaviors. This study is part of a wider investigation into context-sensitive approaches to soundscape design as it pertains to safety, where music and/or sounds can contribute to well-being, and to the fostering of social cohesion. It is a follow on project in the Sounding Brighton series, designed and initiated by Brighton and Hove City Council and the Noise Abatement Society [6]. For this reason, whereas many experiments use aversive sounds to try to repel people [7, 8], our experiment made use of non-aversive, context-appropriate sounds in an attempt to test the scope of fostering more "personable" behavior through soundscape management. The fact that soundscape interventions can be used in both exclusive and inclusive ways points to a more general theoretical account of the role of sound in defining the environment [7, 8], whereby sound is considered much more generally as a way of defining social territory.

Fig. 1. Camera view and results of the automated video analysis. The yellow lines illustrate the output of the tracking algorithm The red dots, resulting of the bobbing analysis process, indicate frames where people are estimated to start a new stride/footstep.

It is now firmly established that one of the more basic behavioral regulations that music can induce is "entrainment", where people will synchronize their activity to an external rhythm. Relevant to the current study are the potential effects of music intervention in the urban environment associated with the previously documented, pro-social effects of synchronized movement to music. Evidence suggests that synchronized movement enhances the tendency in people to perceive others as "more like self", resulting in enhanced compassion [9] and results in improved cooperative ability [10]. This has been shown to occur not just in adults, but also in children as young as four years old [11].

Approach and contributions. Using video to automatically extract gait parameters is a promising approach to evaluate the influence of music in the way people walk. Firstly, because it is non-intrusive: people are not required to wear any sensor,

nor to walk along a pre-specified path. Secondly, because this methodology allows gathering a substantial number of samples, which might be difficult to obtain in lab settings.

Existing methods for gait analysis from video typically require foreground segmentation [12], side cameras [13] or more complex articulated human body analysis [14]. Recent approaches [15] exploit the correlation of head motion and walking motion; nonetheless they mostly rely on video captured in lab settings.

In this paper we propose a gait characterization approach based on walking frequency and phase. Frequency is a clue to the walking speed, whereas phase is a clue for entrainment analysis. These parameters are estimated from the upper-body oscillatory motion resulting from walking: in this work, we use the term bobbing for this motion. Although similar in spirit to [15], our approach is shown to work in surveillance settings rather than in the lab, without foreground segmentation, and can be applied to multiple people at the same time.

Paper plan. Section 2 describes the experimental set-up. Section 3 present the walking analysis algorithm. Finally, section 4 presents preliminary results obtained from both annotated and automatically extracted data.

2. EXPERIMENTAL SETUP

Our analysis is conducted in a tunnel (located in Brighton, UK) that separates a busy nightclub area from the beach. We installed a sound system with a wall-mounted loudspeaker in the northern and southern ends of the tunnel, and three CCTV cameras as per Figure 2. Along the floor we placed white stripes made of durable duct tape; these occurred every 210 cm. We focused the video analysis on camera 1, marked with an X in Fig. 2, which camera field of view is shown Fig 1.

To test the capacity of music to induce entrainment and modulation of walking speed at varying tempos, we took three instrumental pieces of music from contrasting genres (classical, swing jazz, and ambient electronica) and digitally adjusted their tempos (without altering the pitch), so that each were presented in a 106 bpm (a pace slightly faster than a normal walking, but still easy to walk in step with) version and a version that was 10% faster. This resulted in set of 6 music conditions, complemented by a silence condition. Note that the above procedure enabled us to unambiguously control for the change in tempo by keeping the style constant; an approach that differs substantially from the Milliman experiments, where the fast tempo vs slow tempo music were not identical music pieces differing only in tempo, but were also distinct in character and style. Another difference, when compared to the Milliman study, is that there was no primary purpose other than traversing the length of the tunnel.

3. VIDEO PROCESSING

To perform automatic gait characterization in a surveillance setting, we propose an original approach which consists in the following steps:

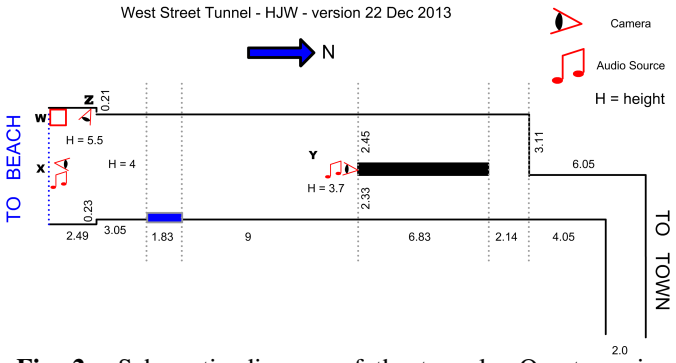

Fig. 2. Schematic diagram of the tunnel. Our team installed several electronic units: W. A locked cubby with 240 VAC mains power, housing the digital video recorder, power amp and music presentation system. The following locations are marked on the diagram: X. Camera 1 (North-facing) + speaker. Y. Camera 3 (South-facing) + speaker. Z. Camera 2 (Southeast-facing) + microphone for ambient noise.

Human detection: we use the deformable parts model (DPM) detector proposed in [16]. Specifically, we use a single mixture trained on full-bodied pictures of people [17].

Tracking by detection: Detections across frames are associated so as to form tracks identifying the walking trajectories of people. A brief description of the tracking approach is provided in Section 3.1.

Motion Estimation. Estimating bobbing could be conducted by analyzing the oscillating sequence of position of body parts like the head. However, this is highly dependent on localization accuracy, which at that resolution can be easily affected by self-occlusion or the presence of texture or people behind the body. Instead, motion, that contains similar bobbing information, is less affected by such inaccuracies since computing motion on different support regions around a given body part produces similar estimates, esp. when a robust estimation method is used. Thus, in this paper, we rely on a robust multi-resolution motion estimation method [18] to estimate an affine motion model using as support region each of the detected body parts of a given human detection.

Bobbing estimation: Using the available detections and visual motion for each track, we estimate bobbing. Details are provided in Section 3.2

3.1. Tracking approach

We follow the method of [19, 20], in which multi-object tracking is formulated as a labeling problem. Namely, given a set of detections $R = \{r_i\}_{i=1:N_r}$ within a video sequence, the aim is to assign an identity label to each of them, so that all detections of the same object have the same label. In other words, the goal is to obtain the label field $L = \{l_i\}_{i=1:N_r}$ such that when detections r_i and r_j represent the same object, then $l_i = l_j$, and $l_i \neq l_j$ otherwise. To that end, we extract for each detection r_i its pixel position X_i and its

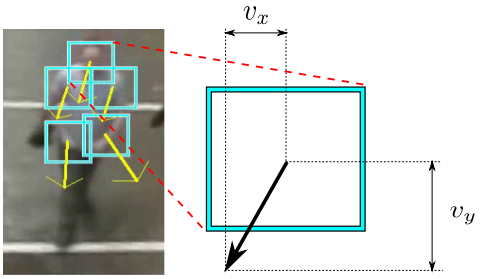

Fig. 3. Parts' estimated motions capture bobbing information. Bounding boxes represent the detected upper body parts and arrows show the estimated motion.

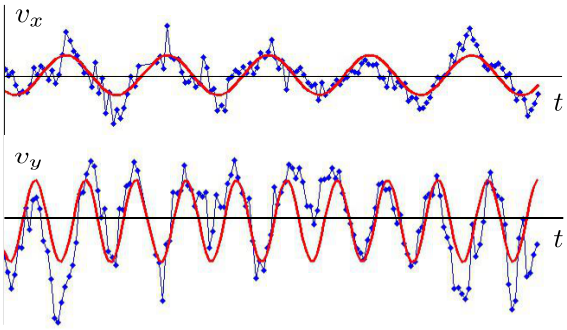

Fig. 4. Sinusoid Fitting. Dotted blue line: time series obtained from head motion vectors, after fitting first order polynomial. Solid red line: fitted sinusoid. Note that the DC component has been removed for visualization purposes.

multi-resolution color histogram h_i, as well as its time of occurrence t_i. These descriptors are used to measure pairwise similarities and dissimilarities between detections. The labeling task is cast into a CRF formulation [19, 20], where we directly model the posterior probability of the label field given all the observations.

3.2. Bobbing estimation

Walking locomotion generates an "up and down" and "left to right" motion in human subjects. In this work, bobbing refers to this motion. Our hypothesis is that this motion can be characterized by sinusoidal functions whose frequency and phase are correlated with the walking speed and phase.

For each track, we rely on the set of motion vectors $\mathbf{v}_t^i$ to estimate a pair of sinusoidal functions modeling bobbing:

$$\begin{aligned} f_x(t, \Lambda_x) &= A_x + B_x \sin(2\pi C_x t + D_x) \\ f_y(t, \Lambda_y) &= A_y + B_y \sin(2\pi C_y t + D_y) \end{aligned} \quad (1)$$

where coefficients $\Lambda = \{A, B, C, D\}$ are the parameters (offset, amplitude, frequency and phase) that define the bobbing.

Although any detected part could be used, in this work we limit ourselves to the head region. Let us assume that given a time interval $[t - T_b + 1, t]$, we have a set of head motion estimates $\mathbf{v}_x = \{v_x(t_0), \ldots, v_x(t_n), \ldots, v_x(t_N - 1)\}$, $\mathbf{v}_y = \{v_y(t_0), \ldots, v_y(t_n), \ldots, v_y(t_N - 1)\}$, where $t_n \in [t - T_b + 1, t]$. Note that we can have missing samples, i.e., N is lower or equal than the maximum number of samples that can be observed in the interval T_b . The proposed bobbing estimation method treats the x and y components independently.

Sinusoid fitting. For each of the time series of a visual motion component, the method proceeds as follows. First, we fit a first order (linear) polynomial to the time series and then subtracted it from the original time series. This step effectively compensates perspective issues of motion estimation in the image plane. Alternatively, we tried adding the first order polynomial as part of the sinuoidal bobbing model of Eq. 1. However, in practice, most of the fitting error is then due to the main slope of the time series, and the fitting process then result in poor estimates of the sinusoid parameters.

Second, we fit the sinusoid model of Eq. 1 to the corrected data using non-linear least squares [21]. Due to noise, estimation errors and bad initial guesses, such optimization might get stuck in local minima. Fortunately, prior knowledge on the average walking speed of humans can be used to robustly fit the model Λ on the data $\mathbf{v}_x, \mathbf{v}_y$. Note that the x component motion frequency is close to 1Hz whereas the vertical component one is typically closer to 2Hz. We can thus initialize C with the aforementioned values to conduct and initial optimization. To improve robustness, we reduced this initial guess through a geometric progression of ratio $\beta = 0.9$, and performed an optimization run for each of the obtained value. At the end of this process (we typically use 10 runs), we keep as estimates the parameters with minimum fitting error.

Foot strike estimation. We formulate the hypothesis that foot strikes occur at the local maxima and minima of the $f_x(t, \Lambda_x)$ function. Accordingly these instants are defined as:

$$\hat{t}_j = \{t_j \mid \frac{d}{dt} f_x(t, \Lambda_x) = 0\}. \quad (2)$$

4. EXPERIMENTATION

We ran a first round of data collection and annotation and we thus provide here the preliminary results that we obtain.

4.1. Collected dataset

To conduct the analysis, a first dataset was collected. It consisted of 47 short video clips (less than 2 minutes) of pedestrian traffic through the tunnel, gathered with motion-sensitive CCTV recordings (see Fig. 2). To study entrainment, these clips were selected under two music conditions, using the same music genre under two different tempos.

We applied automated footfall data tracking to this dataset, and kept only the tracks of pedestrians traveling towards the camera, resulting in a total of 109 people track samples. In a second step, we manually annotated these tracks in two ways. First, we noted the frames when a person's body was crossing the 4^{th} and 1^{st} white lines (see Fig.1), allowing us to measure the average speed of the person for crossing the

tunnel. Secondly, we performed the footfall information annotation by noting the frame number corresponding to subject heel strikes, discounting obscured footfalls. We subsequently compared the list of frame numbers tracking footfalls against the automatic data collected upon the same variable.

4.2. Tracking results

When comparing the footfall data from the automated analysis with data from manual analysis, we found that only 11% (12 tracks) were incoherent. All 97 other tracks were properly following the persons and exhibited no ID switches and no fragmentation. These tracks were however not covering the full tunnel length, with missing parts far from the camera (low resolution) and very close (high distortion). The length of the tracks was still sufficient to estimate the bobbing. Two instances of duplicate tracks (2 tracks for one person) were also observed.

4.3. Bobbing analysis

The bobbing estimation allows us to automatically extract foot strike frames. To evaluate this process, we use the heel strike annotations. As these annotation do not contain occluded heel strikes, we augmented them by interpolation to get the missing heel strikes. A total of 1367 heel strikes were present in the 109 people tracks. Most of the inter-strike times are between 12 and 15 frames in the annotations.

The evaluation is done by matching detected foot strikes with annotated heel strikes. Each strike being used at most once. The evaluation accepts a match if the annotation and the detection instants are at most k frames apart. We noted (and compensated for it) that we had a consistent 1 frame difference between the annotation and the automatic detection. This offset can be explained by the fact that the heel hits the ground slightly before the whole foot. With $k = 2$ (allowing an error of 2 frames), we obtain 77.4% of accuracy (precision and recall). With a stricter criteria, $k = 1$, this accuracy drops to 56.4% which can be due partly to imprecise annotations. With $k = 3$, the accuracy reaches 86.7%.

4.4. Entrainment

Research has shown that unconscious entrainment is very widespread. [22] was able to demonstrate that background music of different tempos played in a supermarket significantly modulates customer velocity. Slower tempo music was consistently associated with slower in-store traffic flow and greater total sales volume versus faster tempo music. Similarly, Milliman demonstrated that background music of different tempos altered the "eating time" of restaurant diners. Faster music was shown to produce faster turnover of tables, yet slower tempo was associated with slower eating and more spending at the bar [23]. More recent research has also indicated that music acts as a far more effective means of modulating walking speed than simply providing a background metronome, suggesting that the broader features

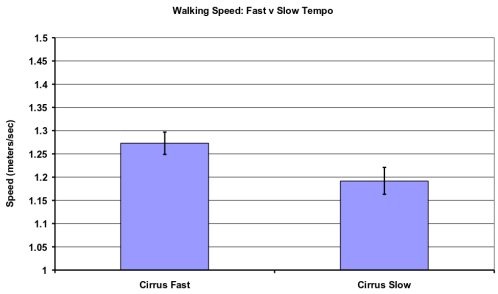

Fig. 5. Difference in walking speed during a fast and slow tempo version of the same musical excerpt.

of music itself, rather than a simple external tempo, play a significant role [24]. In our study, using manual annotations, the same piece of music at two contrasting tempos (all other variables held constant) resulted in significantly different walking velocities. Faster tempo music consistently made people travel quicker through the tunnel (Fig. 5, mean velocity = 1.27m/sec, n = 52) than did slower tempo music (mean velocity = 1.19 m/second, n = 53, unpaired T-test: $P < 0.05$).

These conclusions are particularly intriguing in view of [22] investigation into the effects of ambient/background music on shopper behavior. The difference we observed in walking speed between high and low tempo music is in agreement with Milliman's findings; however, whereas Milliman found higher tempo music was associated with higher customer velocity than no music, we found that both high and low tempo music decreased walking velocity through the tunnel. This divergence might either be ascribed to Milliman's more extreme tempo variations or to our stricter controls for tempo, eliminating style and (other) variables, or alternately it may suggest that music's behavioral effects are highly context sensitive. Furthermore, this general decrease in the speed of those walking through the tunnel when music was deployed might reflect a greater sense of security, and thus a preliminary indicator of the potential for music to support safer environments.

5. CONCLUSION

For the automatic estimation of walking velocity and phase in a soundscape experiment in order to assess walking entrainment to 6 music conditions, this first iteration of our model yields broad correlation (87%) with manual footfall analysis. We find this promising as an approach to automated analysis of large video data sets, gathered passively in the wild with close circuit video cameras. Our investigation will continue with further iterations of the model to accommodate absence of human detections in the tracking [25], finalize the automatic speed estimation from the extracted tracks and improve accuracy, and further data collections for finer music entrainment analysis.

6. ACKNOWLEDGEMENT

We gratefully acknowledge funding from Brighton and Hove City Council and the Noise Abatement Society.

7. REFERENCES

[1] D. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Int. Conf. on Multimodal Interfaces (ICMI)*, 2008.

[2] H. Hung and G. Chittaranjan, "The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game," in *ACM Multimedia, Florence*, 2010.

[3] H. Hung and B. Krose, "Detecting f-formations as dominant sets," in *in ICMI, Alicante*, 2011.

[4] C. Chen and J.-M. Odobez, "We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence*, june 2012.

[5] M. Cristani, A. Pesarin, A. Vinciarelli, and M. Crocco adn V. Murino, "Look at whos talking: Voice activity detection by automated gesture analysis," in *Proc. Workshop on Interactive Human Behavior Analysis in Open or Public Spaces, Amsterdam*, 2011.

[6] L. Lavia, M. Easteal, D. Close, H. Witchel, O. Axelsson, M. Ware, and M. Dixon, "Sounding brighton: Practical approaches towards better soundscapes," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 2012, vol. 2012, pp. 436–444.

[7] Lily E. Hirsch, *Music in American Crime Prevention and Punishment*, University of Michigan Press, 2012.

[8] Harry Witchel, *You are what you hear: How music and territory make us who we are*, Algora Publishing, 2010.

[9] Piercarlo Valdesolo and David DeSteno, "Synchrony and the social tuning of compassion.," *Emotion*, vol. 11, no. 2, pp. 262, 2011.

[10] Piercarlo Valdesolo, Jennifer Ouyang, and David DeSteno, "The rhythm of joint action: Synchrony promotes cooperative ability," *Journal of Experimental Social Psychology*, vol. 46, no. 4, pp. 693–695, 2010.

[11] Sebastian Kirschner and Michael Tomasello, "Joint music making promotes prosocial behavior in 4-year-old children," *Evolution and Human Behavior*, vol. 31, no. 5, pp. 354–364, 2010.

[12] N.V. Boulgouris, D. Hatzinakos, and K.N. Plataniotis, "Gait recognition: a challenging signal processing technology for biometric identification," *Signal Processing Magazine, IEEE*, vol. 22, no. 6, pp. 78–90, Nov 2005.

[13] X. Zhou and B. Bhanu, "Feature fusion of side face and gait for video-based human identification," *Pattern Recognition*, vol. 41, no. 3, pp. 778–795, 2008.

[14] G.W. Taylor, L. Sigal, D.J. Fleet, and G.E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," in *Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 631–638.

[15] Sung Uk Jung and Mark Nixon, "Estimation of 3d head region using gait motion for surveillance video," in *Int. Conf. on Imaging for Crime Detection and Prevention*, 2011.

[16] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627 –1645, sept. 2010.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005.

[18] J. M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Jal of Vis. Comm. and Image Representation*, 1995.

[19] Alexandre Heili and Jean-Marc Odobez, "Parameter estimation and contextual adaptation for a multi-object tracking crf model," in *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2013.

[20] A. Heili, A. Lopez-Mendez, and J.-M. Odobez, "Exploiting long-term connectivity and visual motion in crf-based multi-person tracking," *IEEE Transaction on Image Processing*, 2014.

[21] Jorge J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*, G. A. Watson, Ed., pp. 105–116. Springer, Berlin, 1977.

[22] Ronald E. Milliman, "Using background music to affect the behavior of supermarket shoppers.," *Journal of marketing*, vol. 46, no. 3, 1982.

[23] Ronald E. Milliman, "The influence of background music on the behavior of restaurant patrons," *Journal of consumer research*, pp. 286–289, 1986.

[24] Frederik Styns, Leon van Noorden, Dirk Moelants, and Marc Leman, "Walking on music," *Human movement science*, vol. 26, no. 5, pp. 769–785, 2007.

[25] Jian Yao and Jean-Marc Odobez, "Multi-camera multi-person 3d space tracking with mcmc in surveillance scenarios," in *European Conference on Computer Vision, workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications (ECCV-M2SFA2)*, Marseille, Oct. 2008.