The Sound of Motion: Multimodal horse motion estimation from video and audio

Ci Li[1] Elin Hernlund[2] Hedvig Kjellström[1,3] Silvia Zuffi[4]
[1] KTH, Sweden cil,hedvig@kth.se
[2] SLU, Sweden elin.hernlund@slu.se [3] Silo AI, Sweden
[4] IMATI-CNR, Italy silvia@mi.imati.cnr.it

1. Introduction

In this paper, we investigate multisensory integration for capturing the 3D shape and motion of horses. Multiple sensory channels, like vision, hearing, touch, smell and taste, are the way humans experience the 3D world [24]. Among them, vision is the sense that allows us to perceive the motion and appearance of objects. Model-based computer vision methods assume a model of the object is available. For humans, images or videos are interpreted for 3D human shape and pose estimation [1, 7, 12, 10, 22, 26, 11, 8] through SMPL, a 3D statistical shape model [17].

Hearing complements vision, providing another dimension to perception. Many works studied the correlation between sound and motion, such as estimating upper body movements from piano or violin music [23], or synthesizing plausible gestures for a humanoid virtual agent to accompany natural speech [13]. VOCA [2] uses audio to capture an animated 3D human face, mapping the speech to lip movement for the 3D human face model FLAME [15].

The human brain learns the integration between different senses, which deepening our understanding of the world. Much research simulates sensory data integration, combining video and sound. Owens et al. [19] achieve scene analysis through audio and video. Gao et al. [3] perform speech separation by exploiting visual information from the video. Multisensory integration is the basis of perceptual hallucination, such as hearing the sound of feet stepping on the ground makes us imagine human or animal motion. However, very few studies have exploited the complementary nature of vision and sound for 3D motion estimation, especially for animals, to mimic human perception.

In this work, we investigate multisensory integration to recover 3D motion of horses. Specifically, we hypothesize that audio complements video in the task of 3D shape and pose estimation from videos, and audio helps to estimate more accurate poses. To our best knowledge, our study is the first to estimate the 3D shape and pose of horses by combining video and audio.

Our goal is to recover horse motion. Horses are probably one of the oldest domesticated animals and the most relevant animal for human activities like sports and agriculture. This increases the demand for studying horses through markerless motion capture. Moreover, horse motion reconstruction has a high potential to benefit from the audio cue since horses have hard hooves and strong bodies and make characteristic and significant sounds when moving.

We utilize a 3D articulated shape model of horses, hSMAL [14], which parameterizes the shape and pose of the subject. Two architectures are proposed to explore the additional source of information provided by audio for learning to regress model parameters, exploiting audio at training time alone, and at both training and test time (see Fig. 1). We consider: A *model fusion* setting [5], where, at training time, the audio and video features are passed through a shared regression block to predict 3D pose; this architecture does not require audio at test time. An *early fusion* setting [5], where the audio and video features are concatenated and processed through fully connected layers before entering the 3D pose regression module. In general application, we can expect the audio channel to be absent or very noisy, with the sound of the animal overlapped with different ambient sounds. It would be beneficial to employ the model-fusion network that can exploit audio for training but be applied to a silent video. Our results show that this is possible, thus the complementary use of audio and video provides a useful signal for learning to regress 3D pose.

2. Method

2.1. The hSMAL model

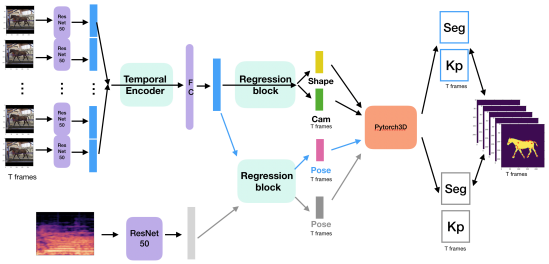
The hSMAL model [14] is a specific SMAL model[28] for horses, with 36 body segments and $1,497$ vertices. The model is learned with the alignment method of [28] and is a template-based 3D model, where the mean template is the average of 37 horse toys scans. The model defines a 3D mesh as a function $M(\beta, \theta)$. β is the shape variable, describing vertex-based deformations with respect to the mean template ; θ is the pose parameter, denoting the relative rotation of each joint to its parent in the predefined kinematic tree in axis angle representation.

2.2. Multimodal model regression

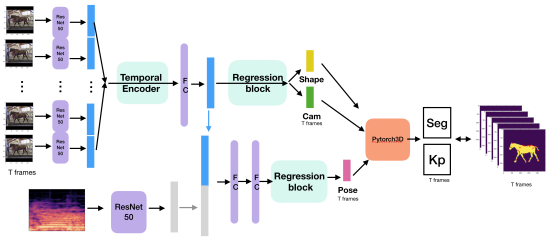
Regression methods for 3D pose estimation usually exploit an inverse rendering approach, given the lack of 3D ground truth. We consider shape, pose and camera estima-

tion, as they all contribute to the rendering of the model.

The visual input $I_{n:n+(t-1)}$ from time step n to $n+(t-1)$ are forwarded to the backbone network for feature extraction. We adopt the temporal encoder from [8] to learn a temporal representation of the video frames. The image features from the backbone are passed through the temporal encoder and a fully-connected layer to get the final feature. The corresponding audio $A_{n:n+(t-1)}$ are transformed to a Mel spectrogram generated by Librosa [18], followed by a backbone network to extract audio features. As in [7], we use an iterative error feedback (IEF) loop, here called regression block, for predicting parameters. We use two regression blocks in our work, named ψ and Φ. Block ψ uses the visual input to predict the weak perspective camera $C_{n:n+(t-1)}$ in t frames and model shape parameters, since we assume the shape of the subject doesn't change during a small period. Block Φ predicts the pose parameters of the model in t frames. We pass the predicted model parameters and the weak perspective cameras into Pytorch3d [20] for rendering silhouettes and 2D keypoints.



(a) Model-fusion.



(b) Early-fusion.

Figure 1: Video-audio fusion frameworks. Video channel (in blue) and audio channel (in gray).

To deal with multimodality data, we choose the model implicit fusion on loss (*model fusion*) and *early fusion* in [5, 6]. We assume visual information is the primary modality, and audio is the auxiliary modality. The network learns the correlation between the main modality and the auxiliary modality during training. During testing, depending on the network architecture, the auxiliary modality might be present or not. In these two frameworks, the camera and the model's shape are predicted through ψ using visual inputs.

Model-fusion network As shown in Figure 1a, the visual and audio features are passed through the shared regression block Φ and output two sets of poses $\theta_{I_{n:n+(t-1)}}$ and $\theta_{A_{n:n+(t-1)}}$, respectively. The two sets of poses are combined with the predicted shape and cameras, passed through Pytorch3d for rendering two sets of silhouette and 2D keypoints. Since two modalities are processed independently, we can handle the situation where one of the modalities is missing in the inference stage.

Early-fusion network As shown in Figure 1b, the visual and audio features are concatenated and passed to two fully connected (FC) layers, inspired by the Temporal Binding Network (TBN) block [9]. The fusion features are then passed through the regression block Φ for pose estimation. We need both video and audio as input during testing.

Loss. Both regression networks are trained in an end-to-end way. We construct the loss function as:

$$L = \sum_{\lambda \in \beta, S_C, Kp_{(\cdot)}, Sil_{(\cdot)}, \theta_{(\cdot)}, S_{\theta_{(\cdot)}}} \omega_\lambda L_\lambda, \qquad (1)$$

where $(\cdot)$ denotes the poses from visual input I or audio A or fusion feature F. L_β and $L_{\theta_{(\cdot)}}$ are the shape and pose prior of the hSMAL model defined in [28, 27, 14]. L_{S_C} is the smooth loss [16, 25] of the predicted camera. $L_{Kp_{(\cdot)}}$, $L_{Sil_{(\cdot)}}$ and $L_{S_{\theta_{(\cdot)}}}$ are the keypoint loss, silhouette loss and smooth loss on the root joint and the rest joints, respectively.

3. Experiments

3.1. The Horse Treadmill Dataset

Our data-driven approach requires observations of horses with videos and audio. The Horse Treadmill Dataset is collected by the University of Zürich [21] [1]. The dataset contains videos, audio and 3D motion capture recordings of ten horse subjects trotting on a treadmill. Three subjects are discarded due to camera calibration problems, leaving seven subjects, with one in white and the rest in the dark (brown or black) colors. Each horse is recorded several times. To our best knowledge, this dataset is the only one that contains synchronized videos, audio and motion capture data.

3.2. Implementation

We randomly pick three horses in dark colors for training and randomly generate nine-frame video clips from the data as input. We create two test datasets: Test Data 1, containing data from the other three horses in dark colors and Test Data 2, containing data from the only white horse. Test Data 2 will pose a much greater challenge to the visual regression than Test Data 1, which allows us to evaluate the contribution of the audio cue in both a case with high-quality and lower-quality visual regression output.

We consider two baselines: a network trained with only visual input, which named "Image-only" Network, and the model fusion network where we only input audio for pose regression, named "Audio-only" Network.

[1] Ethical approval for the collection of this dataset (permission number 51/2013) is granted by The Animal Health and Welfare Commission of the canton of Zurich after evaluating the study protocol, and the horse owners gave informed consent for the inclusion of their animals.

3.3. Results

We evaluate the results with the percentage of correct points (PCK), the intersection of unit (IOU) and the mean per 3D joint position error after rigid alignment with Procrustes analysis (P-MPJPE) [4].

Audio exploration. We explore the network's ability to estimate pose information from audio. To show that the network learns to interpret audio, we test the Audio-only network with original audio and white noise. Table 1 shows that the results with white noise are worse than that with the original audio. These results are consistent in the visual example shown in Figure 2 where the model is in two different views. The model with white noise has a downward head and rigid legs from the side view and an unnatural body tilt from the front view.

Table 1: Quantitative Results with original audio or white noise as test input.

TestDataset		PCK@0.1 ↑	IOU ↑	P-MPJPE ↓
Test Data 1	Original audio	**0.910**	**0.622**	**0.169**
	White noise	0.812	0.507	0.211
Test Data 2	Original audio	**0.872**	**0.611**	**0.147**
	White noise	0.82	0.551	0.181

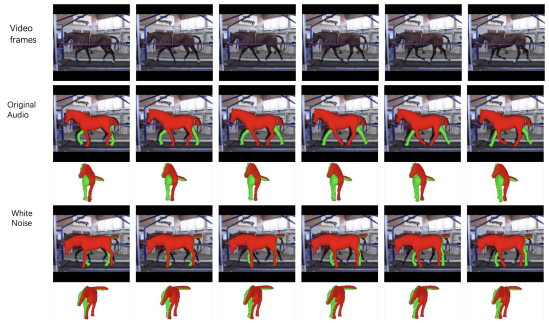


Figure 2: Visual Examples of Audio-only network with original audio or white noise as test input from side and front view. Left body (in red) and right body (in green).

Comparison between different networks. We report all the evaluation criteria for the two baseline networks and the two fusion networks in Table 2.

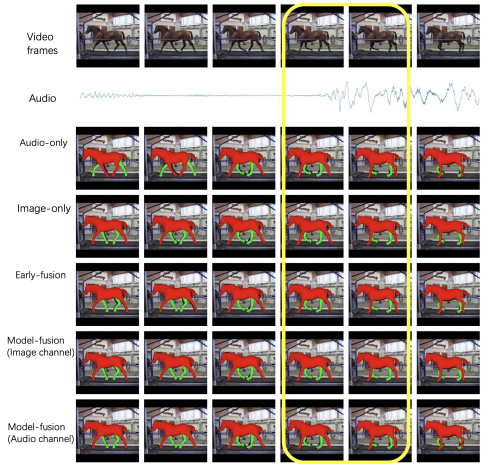
In Test Data 1, the Early-fusion and Image-only networks have similar performance in PCK and IOU. As for P-MPJPE, the Early-fusion network performs better than the Audio-only and Image-only network, as it can learn from both modalities and has the availability of both modalities also at test time. In the Model-fusion network, the audio and image channel perform better than the Audio-only and Image-only networks in P-MPJPE, respectively. As the Model-fusion network effectively benefits from both modalities at training stage and improves the 3D pose prediction from each modality at test time.

In Test Data 2, the network sees a horse in different colors during testing, which brings a more challenging task. The Image-only network performs worse in Test Data 2 than in Test Data 1, while the Audio-only network has mixed results. Even though the visual information varies, the audio is still consistent and preserves the similar information of stepping feet on the ground. The Model-fusion network performs better in all criteria, which shows the benefit.

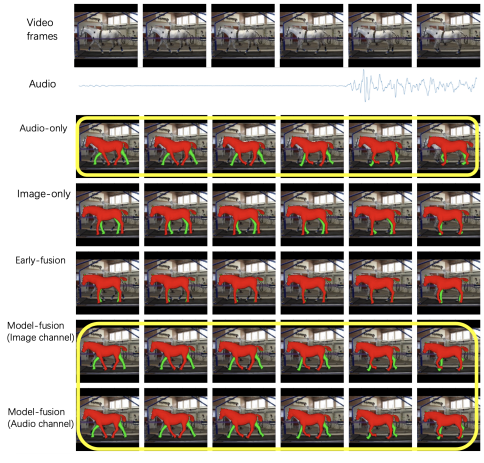
Table 2: Comparison between all networks.

Test Data 1	PCK@0.1 ↑		IOU ↑		P-MPJPE ↓	
	audio	video	audio	video	audio	video
Audio-only	0.910	-	0.622	-	0.169	-
Image-only	-	**0.984**	-	**0.645**	-	0.127
Early-fusion	**0.986**		**0.641**		**0.125**	
Model-fusion	0.919	0.960	0.612	0.622	0.164	**0.125**

Test Data 2	PCK@0.1 ↑		IOU ↑		P-MPJPE ↓	
	audio	video	audio	video	audio	Video
Audio-only	0.872	-	0.6107	-	0.147	-
Image-only	-	0.961	-	0.644	-	0.159
Early-fusion	**0.978**		0.643		0.149	
Model-fusion	0.950	**0.998**	**0.691**	**0.691**	**0.146**	**0.105**



(a) Test Data 1, where the horses in training and testing dataset have dark colors.



(b) Test Data 2, where the testing horse is white.

Figure 3: Visual examples in fusion frameworks.

We show visual examples from different networks for a short sequence when the horse is stepping on the ground. In Test Data 1 (Figure 3a), we can observe that all networks perform similarly. In Test Data 2 (Figure 3b), the Image-only network and Early-fusion network predict rigid legs. The difference of visual information between training and test in the case of the white horse compromises the ability of the Early-fusion network. The Audio-only network and the Model-fusion network can still predict reasonable poses.

When the network can access two modalities during training and testing, we find that the Early-fusion network is more suitable since it can obtain more information from multiple modalities. However, when the main modality (visual information) contains noise or is different from the training dataset, the Model-fusion network could be a better choice since the auxiliary modality can compensate for the main modality.

4. Conclusions

In this study, we evaluated the complementary use of audio and video for horse 3D shape and pose regression from monocular videos. We used the hSMAL model and explored a unique dataset with video, audio, and motion capture data. Our experiments show that the complementary use of audio and video helps to improve 3D pose estimation from monocular video.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 1

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, 2019. 1

[3] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1

[4] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 3

[5] J. Han, Z. Zhang, Z. Ren, and B. Schuller. Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *ICASSP*, 2019. 1, 2

[6] J. Han, Z. Zhang, Z. Ren, and B. W. Schuller. Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *TAC*, 2019. 2

[7] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2

[8] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 1, 2

[9] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019. 2

[10] M. Kocabas, N. Athanasiou, and M. J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1

[11] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 1

[12] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1

[13] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *IVA*, 2019. 1

[14] C. Li, N. Ghorbani, S. Broomé, M. Rashid, M. J. Black, E. Hernlund, H. Kjellström, and S. Zuffi. hsmal: Detailed horse shape and pose reconstruction for motion pattern recognition. *arXiv preprint*, 2021. 1, 2

[15] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *Proc. SIGGRAPH Asia*, 2017. 1

[16] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *TOG*, 2014. 2

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 1

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, 2015. 2

[19] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1

[20] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 2

[21] M. Rhodin, E. Persson-Sjödin, A. Egenvall, F. M. Serra Bragança, T. Pfau, L. Roepstorff, M. A. Weishaupt, M. H. Thomsen, P. R. van Weeren, and E. Hernlund. Vertical movement symmetry of the withers in horses with induced forelimb and hindlimb lameness at trot. *Equine Vet. Journal*, 2018. 2

[22] N. Rueegg, C. Lassner, M. J. Black, and K. Schindler. Chained representation cycling: Learning to estimate 3d human pose and shape by cycling between representations. In *AAAI-20*, 2020. 1

[23] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to body dynamics. In *CVPR*, 2018. 1

[24] T. R. Stanton and C. Spence. The influence of auditory cues on bodily and movement perception. *Frontiers in psychology*, 2020. 1

[25] S. Zhang, Y. Zhang, F. Bogo, P. Marc, and S. Tang. Learning motion priors for 4d human body capture in 3d scenes. In *ICCV*, 2021. 2

[26] T. Zhang, B. Huang, and Y. Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 1

[27] S. Zuffi, A. Kanazawa, and M. J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018. 2

[28] S. Zuffi, A. Kanazawa, D. W. Jacobs, and M. J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017. 1, 2