# RGB-D video-based individual identification of dairy cows using gait and texture analyses

Fumio Okura[a], Saya Ikuma[a], Yasushi Makihara[a], Daigo Muramatsu[a], Ken Nakada[b], Yasushi Yagi[a]

[a] *Department of Intelligent Media, The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan*

[b] *Department of Veterinary Medicine, School of Veterinary Medicine, Rakuno Gakuen University, 582 Bunkyodai-Midorimachi, Ebetsu, Hokkaido 069-8501, Japan*

**Abstract**

The growth of computer vision technology can enable the automatic assessment of dairy cow health, for instance, the detection of lameness. To monitor the health condition of each cow, it is necessary to identify individual cows automatically. Tags using microchips, which are attached to the cow's body, have been employed for the automatic identification of cows. However, tagging requires a substantial amount of effort from dairy farmers as well as induces stress on the cows because of the body-mounted devices. A method for cow identification based on three-dimensional video analysis using RGB-D cameras, which capture images with RGB color information as well subject distance from the camera, is proposed. Cameras are mostly maintenance-free, do not contact the cow's body, and have high compatibility with existing vision-based health monitoring systems. Using RGB-D videos of walking cows, a unified approach using two complementary features for identification, gait (i.e., walking style) and texture (i.e., markings), is developed.

*Keywords:* Dairy cow, Animal identification, RGB-D video, Gait analysis, Texture analysis

## 1. Introduction

Monitoring the health condition of dairy cows is an essential task in dairy farming. Dairy farmers and veterinarians traditionally assess health by manual observation. However, the condition of every cow is not often observed every day because this requires significant time and effort. The resulting lack of daily health management is a major contributor to economic losses.

Thus far, several automatic (or semi-automatic) systems that observe health conditions, such as milking robots that monitor milk quality, support dairy farmers. Computer vision technology, which enables the non-contact observation of cows, has recently attracted attention for health monitoring purposes. Using two-dimensional (2D) cameras that capture RGB color or grayscale intensity images, several studies have aimed to detect lame cows by estimating locomotion scores, which are measured using back shape and gait analysis (Schlageter-Tello et al., 2014; Song et al., 2008; Poursaberi et al., 2010; Pluk et al., 2012; Viazzi et al., 2013). Moreover, Tasdemir et al. (2011) proposed a vision-based estimation of body weight and a body condition score.

Because of the recent popularization of RGB-D cameras, which capture images of RGB color as well as depth (distance), as shown in Figure 1, three-dimensional (3D) imaging has the potential to be the next-generation standard for health monitoring and analysis. Three-dimensional imaging can overcome the restrictions of sensor position.

(a) RGB image. (b) Depth image.

Figure 1: RGB-D image captured from an RGB-D camera mounted in a cowhouse. For the depth image, large depth values (i.e., longer distances from the camera) are shown as brighter pixel values. Missing depth values (due to a weak laser pulse reflection) are shown in black. Note that the aspect ratio and field of view of RGB and depth images are different (i.e., original resolution for RGB image was $1920 \times 1080$ pixels, while the depth image was $512 \times 424$ pixels); a pre-calibrated coordinate mapping between the depth and RGB information from the RGB-D camera can be obtained.

For example, most 2D vision-based lameness detection utilizes side views to extract the back shape while 3D imaging is freed from this limitation and thus even top views can be used for back shape analysis such as in Viazzi et al. (2014), Van Hertem et al. (2014), and Van Hertem et al. (2016). Kuzuhara et al. (2015) established a relationship between body condition scores and 3D shapes extracted from 3D images. In addition, Salau et al. (2015) investigated the characteristics of noise in 3D measurements with respect to the amount of reflection on a cow's coat.

An important requirement for determining trends in

the health condition of every cow in a cowhouse is the automatic identification of individual cows. Dairy cows are traditionally identified using earmarks. One of the oldest electronic cow identification systems attached small transmitters on cows (Bridle, 1976). Recent tags that include integrated circuit (IC) chips enable individual information to be observed by a hand-held terminal or by radio frequency identifier (RFID) sensors (Geng et al., 2009). Identification systems using IC chips and readers are currently implemented in commercial cowhouses such as in automatic milking robots. However, tag-based identification often causes dairy farmers a substantial amount of work in attaching and maintaining the tags for each cow as well as inducing stress on the cows because of the body-mounted devices.

The recent growth in RGB-D vision-based cow health management studies shows that it is valuable to develop cow identification methods that are suitable for RGB-D image analysis. When RGB-D imaging is able to realize vision-based identification, together with health management algorithms, dairy cows will no longer need to have anything attached to their bodies. Thus far, vision-based non-contact identification using 2D cameras has been studied. Mimura et al. (2008) proposed an algorithm for cow identification using RGB color videos of walking cows. This approach compares videos using the motion correlation of textures. Similarly, identification based on texture matching has been studied for other animals (Petersen, 1972; Kelly, 2001). However, it is often difficult to apply these approaches in a practical environment; for example, the appearances of cows change according to the path they take on a wide pathway. In addition, it is difficult to utilize texture-based methods in a dark environment or for breeds with a textureless coat (e.g., Jersey cows).

We therefore propose a new approach that identifies cows based on RGB-D imaging and unifies two features acquired from RGB-D images: gait and texture. The gait- and texture-based features complement each other in vision-based cow identification tasks, and the proposed approach automatically unifies the two identification cues.

Matching textures (i.e., markings) on a cow's coat is an intuitive approach for identifying cows (especially Holstein cows). Texture matching, which is a key problem in computer vision, is employed for a broad range of applications such as object recognition. One of the simplest texture matching methods is the template matching algorithm (Lewis, 1995) which matches two textures based on the subtraction of their image regions. To achieve robustness to changes in illumination and view angle for the target object, local image features (Tuytelaars and Mikolajczyk, 2008) have been designed. A typical example of these features is the scale-invariant feature transform (SIFT) proposed in Lowe (2004). SIFT is based on the histograms of local intensity gradients around a point on an image converted to a scale- and rotation-invariant form. Local features calculated from multiple points on an object are occasionally combined to describe a feature of the object using texton (Leung and Malik, 2001) or bag-of-features approaches (Sivic and Zisserman, 2003). One fundamental drawback when matching local features is mismatching; i.e., if similar features exist at many locations, the local features may match to the wrong points. This problem is mitigated in this work by utilizing the 3D shapes of cows aligned between two RGB-D sequences to limit the search domain for the matching of local features.

Texture matching is affected by appearance conditions such as dark environments, textureless cows (e.g., Jersey cows), and dirt on the cow's coat. Therefore, we also propose another identification algorithm that utilizes the change in body shape while walking (i.e., gait appearance), which is robust to changes in illumination and texture. Gait analysis has been actively studied for vision-based human identification in the biometrics domain (Nixon et al., 2006). Compared with other human identification cues (e.g., faces and fingerprints), gait identification can be employed without the cooperation of the subjects who do not need to perform actions such as facing a camera or putting their finger onto a device to be identified. Thanks to its suitability for use with surveillance cameras, gait identification systems are occasionally employed in forensics applications (Bouchrika et al., 2011; Iwama et al., 2013). Using machine learning and large-scale gait databases that contain the images of walking people (Sarkar et al., 2005; Makihara et al., 2012; Iwama et al., 2012), an identification accuracy of over 95% can now be achieved for over 4,000 people (El-Alfy et al., 2014). Using the input from RGB-D cameras, Nakajima et al. (2013) proposed a 3D gait feature. We firmly believe that gait identification is a promising approach for the vision-based identification of individual cows, who never cooperate by facing a camera nor putting their noses onto a device. One known problem with gait identification is the difficulty of handling appearance changes caused by different walking paths, although a few studies have tackled this problem using a large database (e.g., Muramatsu et al. (2015)). We overcome this problem using a 3D shape alignment process. This process is also utilized for texture identification, which is employed to compensate for appearance changes in the gait identification.

**Contributions.** The main contribution of this study is the development of a cow identification algorithm for practical environments based on RGB-D video sequences. The key problems and approaches for realizing practical identification can be summarized as follows:

- Illumination changes:
  the unification of two complementary features (gait and texture) enable identification under various illuminations (i.e., day and night).
- Various walking paths:
  the 3D alignment method mitigates the negative effect of appearance differences due to a variety of walking paths.


Figure 2: RGB-D image capture system.

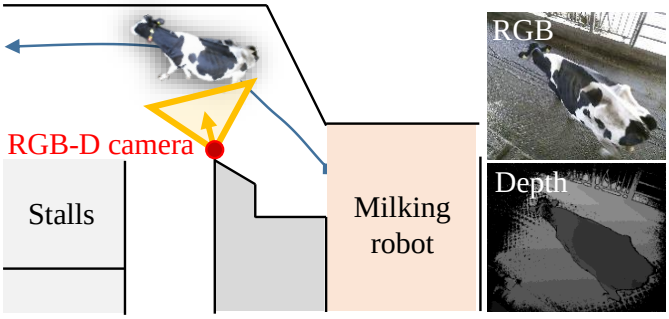



Figure 3: Sensor setting for the experiment.

- Variety of cows:
  gait identification is suitable for cows with textureless or dirty coats.

## 2. Materials and methods

### *2.1. Ethics in animal care*

This study was conducted according to the guidelines issued by the Animal Care and Use Committee of Rakuno Gakuen University. The dataset construction and all other procedures in this study were carried out the without restraining the cows. This study does not require approval from the Animal Care and Use Committee according to the clause in the guidelines (Article 28) that excludes ecology monitoring or livestock sciences from requiring approval.

### *2.2. RGB-D dataset construction*

An automatic capture system for images of walking cows using a Kinect v2 (Microsoft Corporation) RGB-D camera in a water- and dust-proofed box that acquires RGB and depth images at 15 frames per second was developed, as shown in Figure 2. The image capture system detects incoming moving objects using real-time foreground

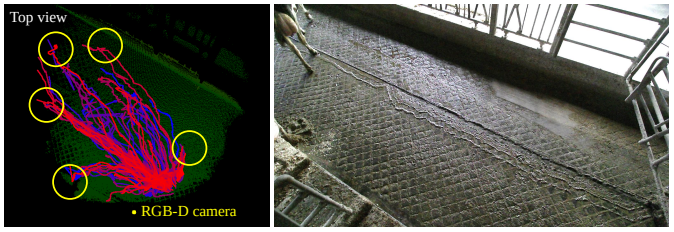



Figure 4: Variation of walking paths in the experimental environment.

extraction, which automatically starts the image capture. The system was installed in an experimental cowhouse in Rakuno Gakuen University, Japan; the RGB-D dataset was created from sequences captured 24 hours per day over a month (from Sep. 28 to Oct. 27, 2015). The camera was attached to a pole at approximately 3.0 m above the ground so that the sensor captured a diagonal view of a pathway which cows walked down from the exit of a milking robot, as illustrated in Figure 3.

The development of an automatic capturing system that only captures walking cows is an important future direction. The video analysis of "walking" cows is beneficial to practical cow monitoring systems such as those designed to assess the back shape of the cow during walking for lameness detection (e.g. Van Hertem et al. (2014)).

The camera occasionally captured incoming objects other than cows, as well as cows that did not always walk smoothly (i.e., cows that occasionally stop). Since gait (and texture) features were selected in this study for cow identification, the sequences were manually selected to meet the following requirements: 1) a cow is captured, and 2) it walks smoothly. The resulting dataset consists of 523 RGB-D image sequences of 16 individual Holstein cows. Note that all individuals were image captured multiple times. To evaluate the identification performance of the system, the individual identification ground truths (i.e., the true identification labels) were provided by the milking robot for all sequences in the dataset. Figure 4 shows the trajectories of the center of gravity of cows from a part of our dataset. The top-view trajectory image (the left image in Figure 4) indicates that the dataset includes a variety of walking paths where cows walk mostly straight but directed at several different areas.

### *2.3. Evaluation methodology*

The basic flow of an individual identification algorithm is as follows. Given a query video sequence (referred to as a *probe* sequence in biometric authentication literature), the identification algorithm searches for candidates of the same individual from the sequences in a previously created dataset (referred to as the *gallery* dataset) using a dissimilarity measure between the probe and gallery sequences.

The identification (i.e., one-to-many matching) accuracy of the proposed identification algorithms was evaluated in this study (the algorithms are detailed in the

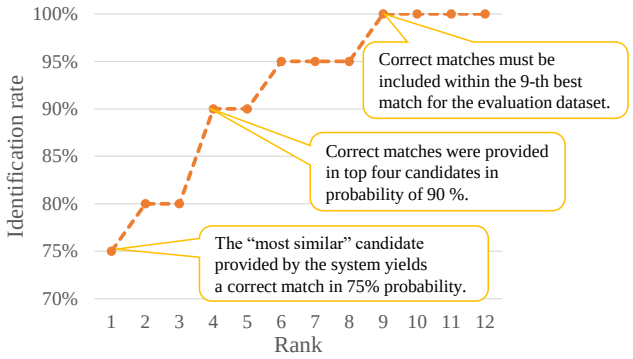



Figure 5: Example of CMC curves, which are utilized for evaluating the identification performance.

following sections). The dataset was divided into two categories: gallery and probe sequences. Sequences captured in the first week (Sep. 28 to Oct. 7; 176 sequences) were designated as the gallery dataset, while the remaining (360) sequences were used as probe sequences. This setting is a simulation of a practical identification scenario, where it takes one week to register the cows in the identification (gallery) dataset.

The identification accuracy was evaluated using cumulative match characteristic (CMC) curves (e.g., Phillips et al. (2000)). CMC curves, which are widely used for evaluating identification performance in biometrics research, show the probabilities of correct matches in the top k-rank matches. Figure 5 shows a typical CMC curve. In the figure, the rank-1 identification accuracy is 75%; i.e., the gallery sequences that yield the smallest dissimilarity for each probe sequence were the correct individual matches with a probability of 75%.

### *2.4. Overview of identification systems*

This paper proposes two vision-based approaches for cow identification: gait and texture analysis. We use images from an RGB-D camera, which has recently become commonly used in dairy research.

Cow identification using the local image features on the cows' coats is proposed. However, this approach is unsuitable for textureless cows, dark environments, or dirty coats. Thus, a gait feature based on a series of walking cow body shape images, which is observed from the depth images, is also proposed. For both features, the appearance changes owing to the difference in appearance when cows walk along different paths are problematic. Therefore, preprocessing to align the 3D shape sequences was employed, and then gait and texture features were calculated using the aligned 3D shapes.

### *2.5. Preprocessing*

Given a depth sequence of a walking cow captured by an RGB-D camera, the proposed system automatically generates an aligned 3D point cloud (i.e., a set of 3D points) sequence of the cow.

#### *2.5.1. Extraction of 3D cow models*

The proposed system automatically extracts the walking cow region based on background subtraction. Using a depth image captured without any cows, the moving object (referred to as the foreground) is extracted using a subtraction approach based on Bayes' theorem. Basically, background subtraction using RGB images in a real environment is a challenging task because of changes in illumination, but by leveraging the depth information, a more accurate foreground region can be acquired.

A statistical model of the background was generated from a depth image sequence without any cows. For each pixel, the following parameters are calculated:

- $P(O=1|X=B)$: probability that depth values are observed during the background sequence
- $\mu$: average depth of the pixel within the sequence
- $\sigma^2$: variance of the depth of the pixel within the sequence

Here, $O \in \{0,1\}$ indicates whether the depth pixel is observed ($O=1$) or missing owing to a weak reflection of the laser pulse ($O=0$). Moreover, $X \in \{F,B\}$ describes whether the observed depth pixel belongs to the foreground ($X=F$) or background ($X=B$).

When a depth image containing cows is observed, the posterior probability that each pixel in the depth image belongs to the foreground is calculated using the prior background model. According to Bayes' theorem, posterior probability $P(X=F|O=1,d)$ for depth value $d$ is calculated as follows:

$$P(X=F|O=1,d) = \frac{P(d|X=F)P(O=1|X=F)}{P(d|X=B)P(O=1|X=B)+P(d|X=F)P(O=1|X=F)}, \quad (1)$$

$$P(X=F) = P(X=B) = 0.5, \quad (2)$$

$$P(O=1|X=F) = 0.5, \quad (3)$$

$$P(d|X=B) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(d-\mu)^2}{2\sigma^2}\right\}, \quad (4)$$

$$P(d|X=F) = \frac{1}{N_d}, \quad (5)$$

where $N_d$ denotes the maximum depth value (8,000 mm in our experiment). Here, priors for the foreground model were not used because the background statistics were modeled using only background sequences. Since the statistical parameters $\mu$ and $\sigma$ are fixed for the same camera setting, this computation is only required once when setting an RGB-D camera at a certain position. An advantage of this approach is its simplicity; capturing a sequence without cows is only required to prepare the foreground extraction process. If several sequences including foreground objects were used as training sequences, the foreground can also be modeled. However, this requires ground-truth labels

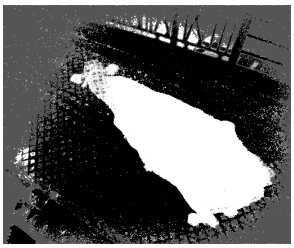


Figure 6: Posterior foreground probability. High probability is indicated by brighter values.

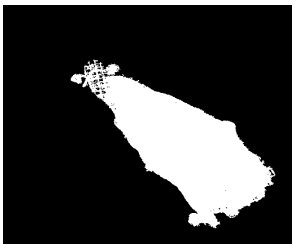


Figure 7: Cow silhouette after thresholding.

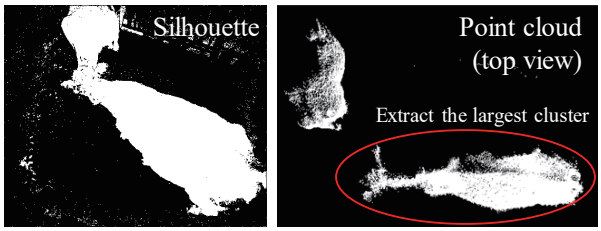



Figure 8: The largest 3D point cluster is extracted for isolating multiple overlapping cows in a silhouette image.

for foreground areas for the training sequences. The labeling of the foreground/background for each frame is quite time-consuming and was thus avoided. The implementation for the foreground extraction in this study is based on a previous gait analysis system using an RGB video sequence that performed well in practice (Makihara et al. (2016)).

The posterior foreground probability $P(X = F|O = 1, d)$ of each pixel was mapped onto an image (see Figure 6) and converted into a silhouette of the cow by thresholding the probability map, as shown in Figure 7. Depth values were converted to 3D pointclouds by placing a 3D point for each depth pixel within the silhouette.

When multiple cows are captured simultaneously, their silhouettes can overlap in a depth frame. Thus, from the 3D pointclouds generated above, the proposed system selects the largest 3D point cluster using Euclidean cluster extraction (Schnabel et al., 2007; Rusu and Cousins, 2011), as shown in Figure 8.

#### *2.5.2. 3D alignment of a cow's shape*

The variation of a cow's walking path is a significant problem in texture and gait analysis. From a camera which is set on a fixed position, the appearance of the texture and body shape changes; accordingly, the dissimilarity computed from the different appearances leads to a large variation. To eliminate changes in appearance due to cows walking along various paths, the position and orientation of point cloud sequences were aligned so that the shape always appears at the same position in 3D space, as shown in Figure9. An iterative closest point (ICP) algorithm (Segal et al., 2009), which is a method for aligning a pair of 3D point clouds for rigid objects, was utilized.

The 3D shapes of walking cows are non-rigid, and adapting the ICP algorithm to this non-rigid case is a practical challenge. To achieve robust matching, the point-to-plane distance (Segal et al., 2009) , which is robust for surface matching including uneven point sampling, was used. The transformation parallel to the ground plane was also restricted. To address the movement of body parts (e.g. legs), a random sample consensus (RANSAC) approach (Fischler and Bolles, 1981) was used to ignore outlier correspondence where the distance is larger than the threshold (0.05 m in our setting).

As shown in Figure 9(a), a point cloud frame was first aligned with its neighboring frames because ICP algorithms occasionally fail to align a pair of point clouds with large distances caused by a large frame gap. Starting from the base frame, which was selected to be the middle frame of the walking sequence, the neighboring frames were aligned so that each point cloud was fitted well to the already-aligned one.

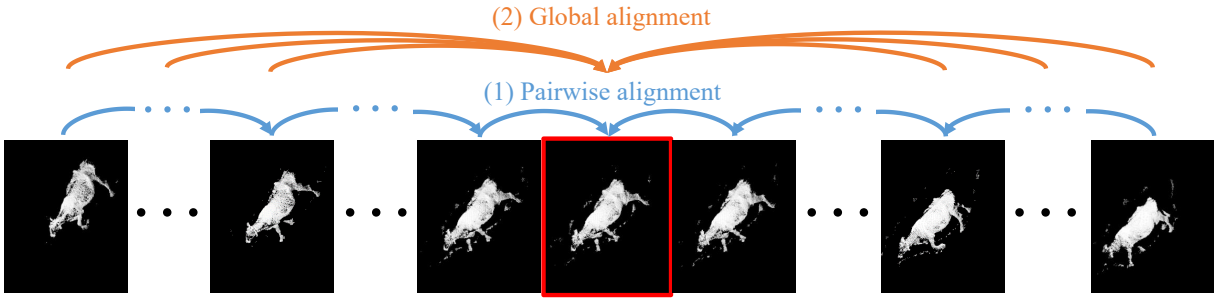
Pairwise alignment accumulates 3D errors, so a global alignment process was performed. Each frame in the sequence was aligned to the base frame using the ICP algorithm. The coordinate system of the 3D space was then converted using the following rules: 1) the first axis of the 3D coordinate system corresponds to the major axis of the cow's body, and is oriented parallel to the ground plane, which can be calculated from the position and orientation of the RGB-D camera. 2) The second axis is perpendicular to the ground plane. 3) The third axis is perpendicular to the other axes. An aligned sequence of point clouds is shown in Figure 9(b), while Figure 10 shows the RANSAC outliers (shown in red) for the final global alignment, and a comparison of combined point clouds before and after the alignment process.

### *2.6. Gait identification*

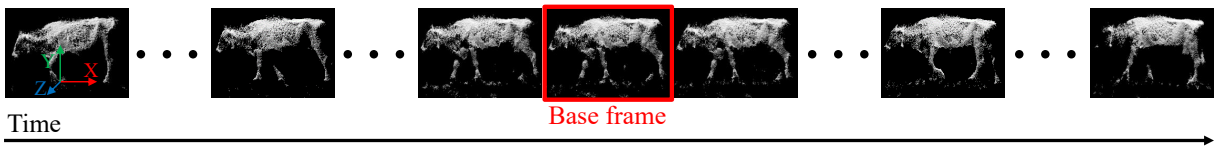
Video-based identification processes are generally performed by comparing features extracted from video sequences. We thus follow the basic pipeline and describe the algorithms for calculating features as well as define the dissimilarity between two sequences. This section describes a cow identification algorithm using a pair of gait features extracted from the aligned 3D pointclouds of gallery and probe sequences. Our gait feature is based on an averaged gait silhouette (Liu and Sarkar, 2004; Han and Bhanu, 2006), which is simple yet known to be an effective feature for the identification of people.

#### *2.6.1. Feature calculation: Averaged silhouette*

The gait feature was computed by averaging aligned silhouettes over one walking period, as illustrated in Figure 12. In the original averaged silhouette method and its extensions (Liu and Sarkar, 2004; Han and Bhanu, 2006),

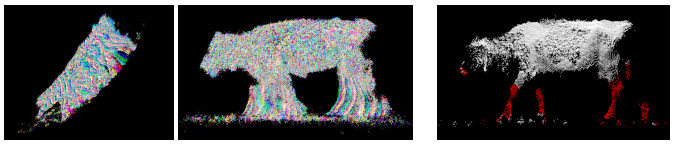



(a) Alignment processes for the 3D points. Neighboring frames are first aligned with each other, and then every frame is aligned to the base frame (the middle frame of the sequence).




(b) Aligned pointclouds. One axis of the 3D space corresponds to the major axis of the cow's body, and another axis is set perpendicular to the ground plane.

Figure 9: 3D shape alignment.



(a) Pointclouds before and after alignment (b) Outliers (shown in red)

Figure 10: 3D alignment result. (a): Point cloud of each frame combined using different colors. (b): Outliers of the ICP algorithm based on RANSAC-based outlier rejection of a frame.

Figure 11: An example of leg occlusion which is not easily identified using side views typically used for gait recognition for humans.

gait features are computed from a 2D silhouette image sequence. Our 3D alignment technique is expected to adapt to the large appearance differences of the sequences.

First, silhouette images observed from the same position relative to the cow's body are virtually generated. Experimentally, the virtual viewpoint is set diagonally upward 50 degrees from the side of the body, which is the approximate average viewpoint of the sensor setting relative to the walking cows. The location of the virtual camera was intended to avoid large missing areas by self-occlusions. For example, legs are sometimes missing due to occlusions as shown in Figure 11, and identification using side views, which is usually used for gait recognition for humans, is not practical. Note that physically capturing side-view sequences in cowhouses is often impractical because the cows usually try to contact the devices and diagonal- or top-views are practical camera settings.

To extract a walking period, the prior knowledge that walking behavior is a periodic motion was utilized. The number of frames belonging to a walking period $N_{gait}$ was calculated by autocorrelation of the silhouette sequence as follows:

$$N_{gait} = \underset{N}{\operatorname{argmax}}\, C(N), \tag{6}$$

where $C(N)$ is the autocorrelation function when the sequence shifts $N$ frames.

$$C(N) = \frac{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n)g(x,y,n+N)}{\sqrt{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n)^2}\sqrt{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n+N)^2}}, \tag{7}$$

$$T(N) = N_{total} - N - 1, \tag{8}$$

where $g(x,y,n)$ describes the silhouette value $\{0,1\}$ at pixel $(x,y)$ in the $n$-th frame and $N_{total}$ is the total number of frames in the sequence.

Image sequences can include multiple walking periods. A gait feature of the $i$-th walking period $E_i(x,y)$ is calculated by averaging the silhouette over one period as follows:

$$E_i(x,y) = \frac{1}{N_{gait}} \left| \sum_{n=iN_{gait}}^{(i+1)N_{gait}-1} g(x,y,n) \right|. \tag{9}$$

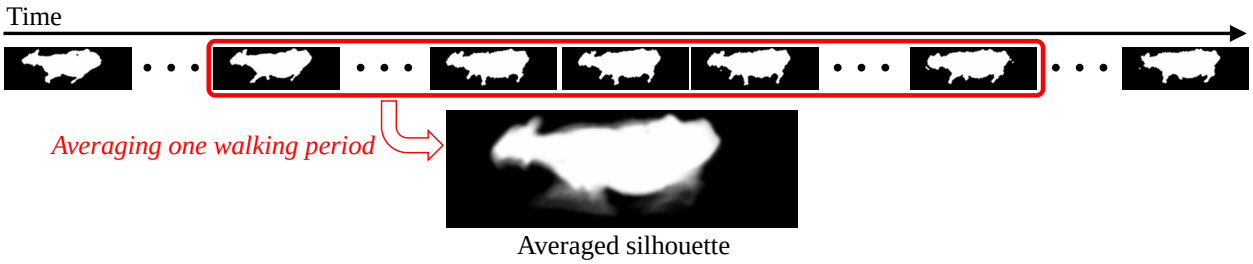



Figure 12: Gait feature generation.

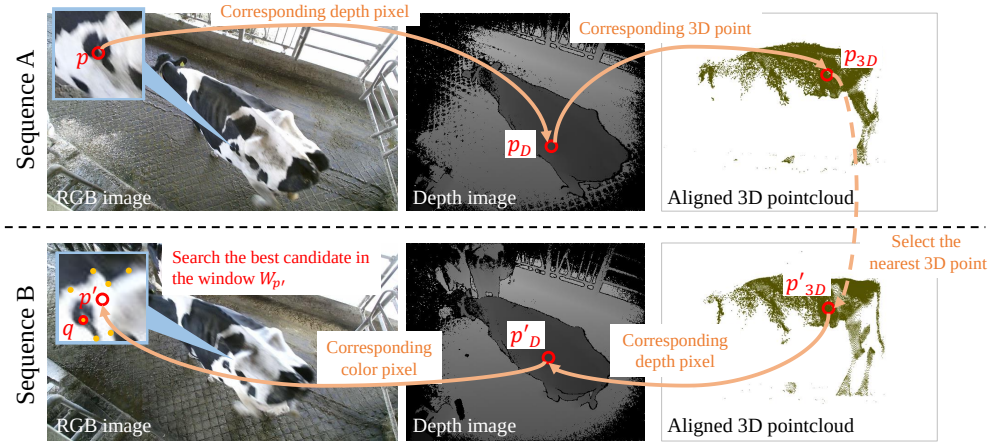



Figure 13: Aligned local matching of SIFT features.

Note that the size of gait features was not *normalized* when the process was included in the generation of gait energy images (GEI), a widely-used gait feature (Han and Bhanu, 2006). The feature used in this work is similar to the original "averaged silhouette" gait feature (Liu and Sarkar, 2004), which does not include the size normalization. Normalization was not employed because metric (scale) information was acquired through depth sensing. One important role of the normalization of GEI is to align the silhouette sequences with the unknown scale factor in RGB videos. Because the scale information is known, the difference in body sizes was represented in our feature.

### *2.6.2. Dissimilarity definition: Euclidean distance*

The dissimilarity between gait features calculated from two sequences is based on the Euclidean distance. Letting the features computed from two sequences be $E_{A_i}(i = 1, 2, ...)$ and $E_{B_j}(j = 1, 2, ...)$, the dissimilarity between the two sequences $D_{gait}$ is defined as the median of the most similar sequence for each walking period.

$$D_{gait} = \mathrm{Median}_i \left[ \min_j \| E_{A_i} - E_{B_j} \|_2 \right], \qquad (10)$$

where $\| \cdot \|_2$ denotes the L2 norm.

## *2.7. Texture identification*

Texture is an intuitive feature that farmers use to identify the individuals of some cow breeds (e.g., Holstein). Several studies have proposed texture-based cow identification, as described in Section 1; however, these methods are difficult to adapt to environments in which viewpoints or walking paths can change. Local image features are occasionally employed for identifying objects. Recent studies report that invariant features based on gradient histograms, such as the SIFT feature (Lowe, 2004), perform robustly against viewpoint, illumination, and shape changes of objects in images. Following this recent trend, as well as for overcoming the disadvantages of local features, an identification approach based on the local matching of SIFT features, where the search domain for the feature matching is limited by the aligned 3D point clouds, is proposed.

The implementation of SIFT feature extraction and dissimilarity calculation use one image per sequence (e.g., the middle frame); however, the algorithm can be extended to employ the whole sequence.

### *2.7.1. Feature calculation: SIFT features*

SIFT is a widely used local feature that is invariant to changes in rotation, scale, and illumination. Although the combination of scale and rotation invariance does not exactly represent affine or perspective transformation in principle, the original SIFT paper (Lowe, 2004) estimated the image correspondence under affine and perspective transforms and performed well if the view difference is

not significant, i.e. if the view difference is sufficiently small, the transformation in local areas can be simulated by the combination of rotation and scale change. This is one reason why SIFT features are often utilized for estimating image correspondences under practical situations (e.g., stereo matching involving viewpoint change).

SIFT features are calculated from gradient histograms around keypoints, which are distinguishable image points extracted near corners or edges. The algorithm automatically extracts SIFT features on RGB images and picks up the features only within the cow's silhouette, which is computed by the process described in Section 2.5.1.

*2.7.2. Dissimilarity definition: Aligned local matching*

SIFT features, which are represented by 128-dimensional feature vectors, are basically matched using the Euclidean distance between the feature vectors. However, just searching for similar keypoints between a pair of images without any restriction causes a large number of mismatches. The aligned 3D point clouds were utilized to limit the search domain and overcome this problem.

Figure 13 illustrates the matching process which searches for keypoint $q \in S_B$ in Sequence B corresponding to keypoint $p \in S_A$ in Sequence A, where $S_A$ and $S_B$ denote sets of keypoints in the sequences. For each keypoint in Sequence A, the corresponding points on depth image $p_D$ and 3D point cloud $p_{3D}$ in the same frame of the sequence were calculated based on the calibration of the RGB-D camera. Candidate point $p'_{3D}$ in Sequence B was then determined to be the nearest point in 3D space. Using the camera calibration, $p'$ was calculated on the RGB image in Sequence B.

The corresponding point $q$ in Sequence B is defined as the keypoint that is the most similar to $p$ within a certain window size centered on $p'$ using the Euclidean distance between SIFT feature vectors as follows:

$$q = \underset{x \in W_{p'}}{\operatorname{argmin}}(\|\boldsymbol{v}(p) - \boldsymbol{v}(x)\|_2), \tag{11}$$

where $x \in W_{p'}$ denotes all SIFT keypoints in a window centered on $p'$, which are candidates for a corresponding point. Moreover, $\boldsymbol{v}(\cdot)$ denotes a feature vector corresponding to a keypoint.

The dissimilarity $D_{tex}$ between two sequences is defined as the average dissimilarity among all pairs of corresponding points. In addition, $D_{tex}$ was set to infinity when the variance of the image intensity over a cow region $\sigma$ is smaller than the threshold $\theta$, because texture-based identification is likely to fail when texture variation is small (e.g., at night or on textureless cows). Hence, $D_{tex}$ is defined as:

$$D_{tex} = \begin{cases} \frac{\sum_{p \in S_A} \|\boldsymbol{v}(p) - \boldsymbol{v}(q)\|_2}{|S_A|} & (\sigma \geq \theta) \\ \infty & (\sigma_i < \theta) \end{cases} \tag{12}$$

where $|S_A|$ denotes the number of keypoints in Sequence A.

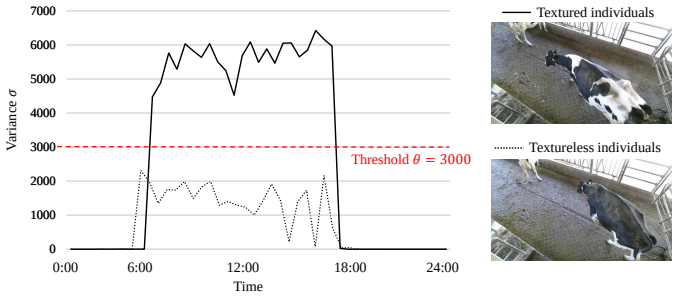



Figure 14: Image variance $\sigma$ for textured and textureless individuals.

Here, the threshold $\theta$ was determined from training sequences so that the images of *textured* cows (individuals clearly including both white and black surfaces) captured in the daytime were judged as textured sequences, and *textureless* individuals were judged as textureless sequences even in the daytime. Figure 14 shows the relationship between the variance $\sigma$ and time of day; $\theta = 3000$ was used in grayscale images whose intensity ranges from 0 to 255.

*2.8. Score-level fusion of two identification cues*

Gait identification and texture identification complement each other in cow identification scenarios in changing environments; hence, it is desirable to automatically combine these approaches. The fusion of multiple (e.g., gait and face) cues has been studied for human identification tasks. Leveraging a machine learning approach on a large dataset constructed under various environments, score-level (Ross et al., 2006) and feature-level (Ross and Govindarajan, 2005) fusion have been shown to significantly improve identification accuracy.

A simple score-level fusion that linearly combines the gait- and texture-based dissimilarities was implemented. Meanwhile, employing texture-based dissimilarity under textureless scenes (computed based on the intensity variance $\sigma$ above) was avoided. Thus, the combined dissimilarity is calculated as:

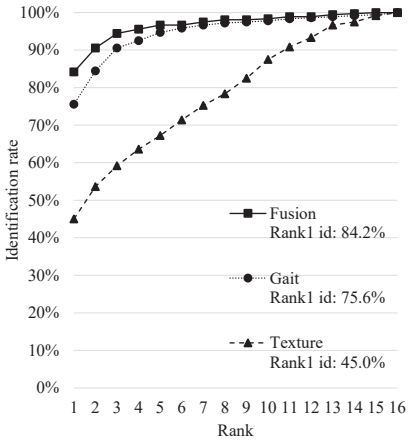
$$D = \begin{cases} \alpha \bar{D}_{tex} + (1-\alpha)\bar{D}_{gait} & (\sigma_i \geq \theta) \\ \bar{D}_{gait} & (\sigma < \theta) \end{cases} \tag{13}$$

where $\bar{D}_{tex}$ and $\bar{D}_{gait}$ denote the normalized dissimilarities, which were calculated as dissimilarities divided by the average of each dissimilarity distribution over the dataset. Note that $\alpha$ is a factor for the linear combination, and $\alpha = 0.8$ was used empirically, heavily weighting the texture cue in the textured environment.

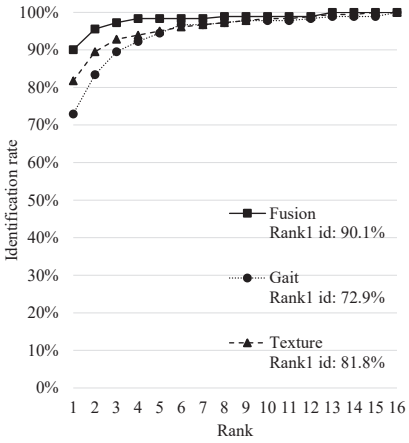
## 3. Results

*3.1. CMC curves of both identification algorithms*

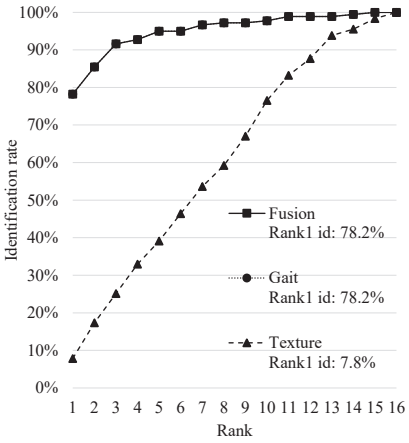
Figure 15 shows CMC curves for the two identification cues as well as their fusion. RGB information a with low signal-to-noise (S/N) ratio in a dark environment (e.g., night), as well as the lack of texture on a cow's coat caused




(a) Results for all probe sequences.




(b) Results for probe sequences captured in scenes with textures ($\sigma \geq \theta$).

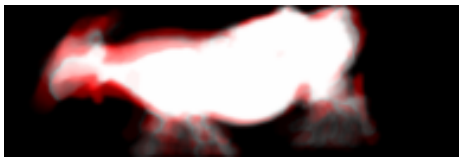



(c) Result for probe sequences captured in textureless scenes ($\sigma < \theta$).

Figure 15: CMC curves of cow identification. The fusion of two identification cues achieves high identification accuracies in every environment. In textureless scenes (night or textureless cows), texture identification gives nearly chance-level results (a rank-1 identification rate of 7.8 %). Note that the dissimilarity of the fusion method is equal to the dissimilarity of gait features in (c); thus, the two CMC curves overlap.

(a) Gallery gait feature.

(b) Probe gait feature.



(c) Difference of gait features shown in red.

Figure 16: Comparison of the gait features of the same individuals where the correct match was not obtained in the first rank.

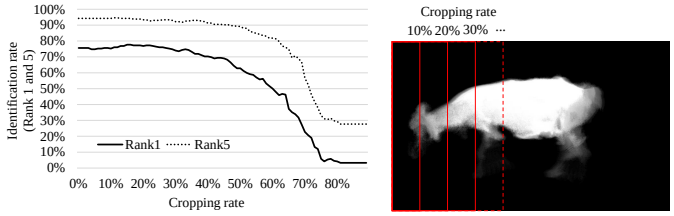



Figure 17: Gait identification result when simply removing head part.

by dirt or the breed (e.g., those with uniformly colored coats), were a source of difficulty for texture identification. As expected, texture identification yielded a high rank-1 identification accuracy (90.1%) in scenes with textures (see Figure 15(b)), while it gives almost chance-level results in textureless scenes such as night scenes or those of cows with textureless coats. In contrast, gait identification using 3D analysis is not affected by scene brightness or the availability of textures because the depth image sensing using laser pulses is less affected by a dark environment. Our hybrid approach (called "Fusion" in Figure 15) achieves the best rank-1 identification accuracy for every environment.

## 4. Discussion

### *4.1. Accuracy of gait identification*

The gait-based cow identification algorithm achieved a 75.6% identification accuracy, whereas human identification accuracy using a similar feature (Han and Bhanu, 2006) was 94.2% on a dataset consisting of over 4,000 persons (Iwama et al., 2012; El-Alfy et al., 2014). Here, we investigate why the accuracy of cow identification was lower than that of human identification.

Figure 16 shows the gait features of the same individuals for which the gait identification failed to find a correct match in rank 1. The difference of the two features (shown red in Figure 16(c)) shows that a large error appears around the head. One reason for the error could be that cows occasionally move their heads while walking, and this behavior is independent of the periodic motion of walking. This is an animal-specific problem that has been generally ignored in human identification tasks.

*Part-based gait identification.* One approach for overcoming the head-sway problem is to ignore the head part during gait identification. Part-based approaches (Felzenszwalb et al., 2010), which utilize features on specific parts on the body, are employed for pedestrian detection in environments in which occlusions occur. To investigate the applicability of this strategy to cow identification, a simple parts-based gait recognition approach ignoring the head

part was implemented. As shown in Figure 17, the head part of cows was removed by cropping a certain amount from the side gait feature. The left graph in Figure 17 illustrates the change of rank-1 and rank-5 identification rates when cropping gait features as a function of the crop area (see the right in the figure). The rank-1 identification rate has been only slightly increased (from 75.6 % to 77.2 %) when cropping almost all parts of the head (i.e., approximately 20 % crop). This indicates that multiple factors affect performance degradation and this needs to be resolved in future work. For further investigation using large-scale datasets, optimization of distances using metric learning (e.g. Makihara et al. (2017)) may be a good direction.

*Multi-modal gait identification.* While we used two modalities, gait and texture, for the individual identification task, employing additional modalities is another strategy for improving the matching accuracy. In this work, unlike usual approaches for creating a widely used gait feature (GEI) (Han and Bhanu, 2006), the size of gait features was not normalized. This is because the scale information was acquired through depth sensing and we wanted to include the difference in body size in our feature. Our approach can be categorized as a simple version of feature-level fusion of two modalities: body size and gait. Since the size and gait information are encoded into the same gait feature, it is difficult to estimate the contribution of each modality to the identification accuracy. However, we believe the impact of size information is limited compared to the gait feature. According to a human gait analysis study by Uddin et al. (2017), rank-1 identification rate using a person's height was only 0.8 %, while the accuracy using the gait feature (GEI) was 89.7 %. For further development, the presence of metric information can provide interesting applications for gait analysis. For instance, metric information such as the length of the legs are promising features in addition to gait.

*Size of the dataset.* Our experiment employed a dataset consisting of 523 sequences of 16 individual cows. In contrast, several studies utilizing 3D videos have been larger in scale (e.g., Van Hertem et al. (2014) employed a dataset consisting of 511 individual cows). Our experimental results demonstrate that our algorithms are useful for relatively small cowhouses with tens of individual cows. However, we plan to collect a dataset including a larger number of individuals and sequences with multiple cowhouses in addition to long-term observation. A large-scale dataset that includes hundreds or thousands of individuals is useful for more reliable evaluation as well as for machine learning algorithms such as silhouette-based gait recognition leveraging deep learning (Takemura et al., 2018).

### *4.2. Toward fully automatic systems*

In this study, we manually selected RGB-D sequences that captured walking cows for input into the proposed system. Manual selection is a major challenge that needs to be eliminated in order to achieve a fully automatic system that operates in practical environments. We plan to implement an automatic detector for walking cows using object detection (e.g., Okuma et al. (2004)) and tracking (e.g., Wu et al. (2013)) techniques.

## 5. Conclusion

This paper described two complementary features, gait and texture, from RGB-D images for cow identification, and an identification method using a fusion of the two identification cues. Leveraging the alignment algorithms of cow shape 3D point clouds that are automatically extracted by a background subtraction of depth images, the gait feature is calculated by averaging the aligned silhouette sequence of a walking cow. The texture dissimilarity between two sequences is calculated as the average dissimilarity of the SIFT features, where the search domain is limited using the 3D alignment information. We implemented a score-level fusion approach to selectively combine the dissimilarities of texture and gait features.

In an experiment using sequences captured in a cowhouse over a month, the accuracy (the rank-1 identification rate) of our unified approach was 84.2 %. This is superior to the individual accuracies of gait identification and texture identification. A future research direction is the development of a fully automated system for detecting walking cows.

Together with cow health monitoring systems using RGB-D cameras (Viazzi et al., 2014; Kuzuhara et al., 2015; Van Hertem et al., 2016), we are confident that RGB-D video-based identification approaches will become a key technology in next-generation dairy farming.

## Acknowledgments

This research was partially supported by JSPS Grants-in-Aid for Scientific Research, Grant Number 17K12715, and JST PRESTO, Grant Number JPMJPR17O3. We would like to thank Editage (www.editage.com) for English language editing.

## References

Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.. On using gait in forensic biometrics. Journal of Forensic Sciences 2011;56(4):882–889.

Bridle, J.E.. Automatic dairy cow identification. Journal of Agricultural Engineering Research 1976;21(1):41–48.

El-Alfy, H., Mitsugami, I., Yagi, Y.. A new gait-based identification method using local gauss maps. In: Proc. Human Gait and Action Analysis in the Wild: Challenges and Applications (in conjunction with ACCV'14). 2014. p. 3–18.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.. Object detection with discriminatively trained part-based models. IEEE Trans on Pattern Analysis and Machine Intelligence 2010;32(9):1627–1645.

Fischler, M.A., Bolles, R.C.. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 1981;24(6):381–395.

Geng, L., Qian, D., Zhao, C.. Cow identification technology system based on radio frequency. Trans of the Chinese Society of Agricultural Engineering 2009;25(5):137–141.

Han, J., Bhanu, B.. Individual recognition using gait energy image. IEEE Trans on Pattern Analysis and Machine Intelligence 2006;28(2):316–322.

Iwama, H., Muramatsu, D., Makihara, Y., Yagi, Y.. Gait verification system for criminal investigation. IPSJ Trans on Computer Vision and Applications 2013;5:163–175.

Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans on Information Forensics and Security 2012;7(5):1511–1521.

Kelly, M.J.. Computer-aided photograph matching in studies using individual identification: an example from Serengeti cheetahs. Journal of Mammalogy 2001;82(2):440–449.

Kuzuhara, Y., Kawamura, K., Yoshitoshi, R., Tamaki, T., Sugai, S., Ikegami, M., Kurokawa, Y., Obitsu, T., Okita, M., Sugino, T., Yasuda, T.. A preliminarily study for predicting body weight and milk properties in lactating Holstein cows using a three-dimensional camera system. Computers and Electronics in Agriculture 2015;111:186–193.

Leung, T., Malik, J.. Representing and recognizing the visual appearance of materials using three-dimensional textons. Int'l Journal of Computer Vision 2001;43(1):29–44.

Lewis, J.P.. Fast template matching. In: Proc. Vision Interface '95. 1995. p. 120–123.

Liu, Z., Sarkar, S.. Simplest representation yet for gait recognition: Averaged silhouette. In: Proc. 17th Int'l Conf. on Pattern Recognition (ICPR'04). volume 4; 2004. p. 211–214.

Lowe, D.G.. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision 2004;60(2):91–110.

Makihara, Y., Kimura, T., Okura, F., Mitsugami, I., Niwa, M., Aoki, C., Suzuki, A., Muramatsu, D., Yagi, Y.. Gait collector: An automatic gait data collection system in conjunction with an experience-based long-run exhibition. In: Proc. of the 8th IAPR International Conference on Biometrics (ICB 2016), Article No. O17. Number O17; 2016. p. 1–8.

Makihara, Y., Mannami, H., Tsuji, A., Hossain, M.A., Sugiura, K., Mori, A., Yagi, Y.. The OU-ISIR gait database comprising the treadmill dataset. IPSJ Trans on Computer Vision and Applications 2012;4:53–62.

Makihara, Y., Suzuki, A., Muramatsu, D., Li, X., Yagi, Y.. Joint intensity and spatial metric learning for robust gait recognition. In: Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'17). 2017. .

Mimura, S., Itoh, K., Kobayashi, T., Takigawa, T., Tajima, A., Sawamura, A., Otsu, N.. The cow gait recognition using CHLAC. In: Proc. ECSIS Symp. on Bio-inspired Learning and Intelligent Systems for Security (BLISS'08). 2008. p. 56–57.

Muramatsu, D., Shiraishi, A., Makihara, Y., Uddin, M.Z., Yagi, Y.. Gait-based person recognition using arbitrary view transformation model. IEEE Trans on Image Processing 2015;24(1):140–154.

Nakajima, H., Mitsugami, I., Yagi, Y.. Depth-based gait feature representation. IPSJ Trans on Computer Vision and Applications 2013;5:94–98.

Nixon, M.S., Tan, T., Chellappa, R.. Human Identification Based on Gait. Springer, 2006.

Okuma, K., Taleghani, A., Freitas, N.d., Little, J.J., Lowe, D.G.. A boosted particle filter: Multitarget detection and tracking. In: Proc. 8th European Conf. on Computer Vision (ECCV'04). 2004. p. 28–39.

Petersen, J.C.B.. An identification system for zebra (Equus burchelli, Gray). African Journal of Ecology 1972;10(1):59–63.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.. The FERET evaluation methodology for face-recognition algorithms. IEEE Trans on Pattern Analysis and Machine Intelligence 2000;22(10):1090–1104.

Pluk, A., Bahr, C., Poursaberi, A., Maertens, W., Van Nuffel, A., Berckmans, D.. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. Journal of Dairy Science 2012;95(4):1738–1748.

Poursaberi, A., Bahr, C., Pluk, A., Van Nuffel, A., Berckmans, D.. Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. Computers and Electronics in Agriculture 2010;74(1):110–119.

Ross, A.A., Govindarajan, R.. Feature level fusion of hand and face biometrics. In: Proc. SPIE Conf. on Biometric Technology for Human Identification II. volume 5779; 2005. p. 196–204.

Ross, A.A., Jain, A.K., Nandakumar, K.. Score level fusion. In: Handbook of Multibiometrics. Springer; 2006. p. 91–142.

Rusu, R.B., Cousins, S.. 3D is here: Point Cloud Library (PCL). In: Proc. 2011 IEEE Int'l Conf. on Robotics and Automation (ICRA'11). 2011. p. 1–4.

Salau, J., Bauer, U., Haas, J.H., Thaller, G., Harms, J., Junge, W.. Quantification of the effects of fur, fur color, and velocity on Time-Of-Flight technology in dairy production. SpringerPlus 2015;4(1):144.

Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W.. The humanID gait challenge problem: Data sets, performance, and analysis. IEEE Trans on Pattern Analysis and Machine Intelligence 2005;27(2):162–177.

Schlageter-Tello, A., Bokkers, E.A., Koerkamp, P.W.G., Van Hertem, T., Viazzi, S., Romanini, C.E., Halachmi, I., Bahr, C., Berckmans, D., Lokhorst, K.. Manual and automatic locomotion scoring systems in dairy cows: A review. Preventive Veterinary Medicine 2014;116(1):12–25.

Schnabel, R., Wahl, R., Klein, R.. Efficient RANSAC for point-cloud shape detection. Computer Graphics Forum 2007;26(2):214–226.

Segal, A., Haehnel, D., Thrun, S.. Generalized-ICP. In: Proc. Robotics: Science and Systems (RSS'09). volume 25; 2009. .

Sivic, J., Zisserman, A.. Video google: A text retrieval approach to object matching in videos. In: Proc. Ninth IEEE Int'l Conf. on Computer Vision (ICCV'03). 2003. p. 1470–1477.

Song, X., Leroy, T., Vranken, E., Maertens, W., Sonck, B., Berckmans, D.. Automatic detection of lameness in dairy cattle—Vision-based trackway analysis in cow's locomotion. Computers and Electronics in Agriculture 2008;64(1):39–44.

Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.. On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Trans on Circuits and Systems for Video Technology 2018;28(1).

Tasdemir, S., Urkmez, A., Inal, S.. Determination of body measurements on the Holstein cows using digital image analysis and estimation of live weight with regression analysis. Computers and Electronics in Agriculture 2011;76(2):189–197.

Tuytelaars, T., Mikolajczyk, K.. Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 2008;3(3):177–280.

Uddin, M.Z., Muramatsu, D., Kimura, T., Makihara, Y., Yagi, Y.. Multiq: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. IPSJ Trans on Computer Vision and Applications 2017;9(1):18.

Van Hertem, T., Bahr, C., Schlageter Tello, A., Viazzi, S., Steensels, M., Romanini, C., Lokhorst, C., Maltz, E., Halachmi, I., Berckmans, D.. Lameness detection in dairy cattle: Single predictor v. multivariate analysis of image-based posture processing and behaviour and performance sensing. Animal 2016;10(9):1–8.

Van Hertem, T., Viazzi, S., Steensels, M., Maltz, E., Antler, A., Alchanatis, V., Schlageter-Tello, A.A., Lokhorst, K., Romanini, E.C., Bahr, C., Beckmans, D., Halachmi, I.. Automatic lameness detection based on consecutive 3D-video recordings. Biosystems Engineering 2014;119:108–116.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Ro-

manini, C., Pluk, A., Halachmi, I., Lokhorst, C., Berckmans, D.. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. Journal of Dairy Science 2013;96(1):257–266.

Viazzi, S., Bahr, C., Van Hertem, T., Schlageter-Tello, A., Romanini, C., Halachmi, I., Lokhorst, C., Berckmans, D.. Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. Computers and Electronics in Agriculture 2014;100:139–147.

Wu, Y., Lim, J., Yang, M.H.. Online object tracking: A benchmark. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'13). 2013. p. 2411–2418.