Depth Sensor-Based Facial and Body Animation Control

Yijun Shen, Jingtian Zhang, Longzhi Yang, and Hubert P. H. Shum

Contents

Abstract

Depth sensors have become one of the most popular means of generating human facial and posture information in the past decade. By coupling a depth camera and computer vision based recognition algorithms, these sensors can detect human facial and body features in real time. Such a breakthrough has fused many new research directions in animation creation and control, which also has opened up new challenges. In this chapter, we explain how depth sensors obtain human facial and body information. We then discuss on the main challenge on depth sensor-based systems, which is the inaccuracy of the obtained data, and explain how the problem is tackled. Finally, we point out the emerging applications in the

Y. Shen (✉) • J. Zhang (✉) • L. Yang (✉) • H.P.H. Shum (✉)
Northumbria University, Newcastle upon Tyne, UK
e-mail: yi.shen@northumbria.ac.uk; jingtian.zhang@northumbria.ac.uk; longzhi.yang@northumbria.ac.uk; hubert.shum@northumbria.ac.uk

B. Müller, S.I. Wolf (eds.), *Handbook of Human Motion*,
DOI 10.1007/978-3-319-30808-1_7-1

field, in which human facial and body feature modeling and understanding is a key research problem.

Keywords
Depth sensors • Kinect • Facial features • Body postures • Reconstruction • Machine learning • Computer animation

Introduction

In the past decade, depth sensors have become a very popular mean of generating character animation. In particular, since these sensors can obtain human facial and body information in real time, it is used heavily in real-time graphics and games. While it is expensive to use human motion to interact with computer applications using traditional motion capture system, depth sensors provide an affordable alternative. Due to the low cost and high robustness of depth sensors, it can be applied in a wide application domain with easy setup. Apart from popular applications such as motion-based gaming, depth sensors are also applied in emerging applications such as virtual reality, sport training, serious games, smart environments, etc.

In order to work with depth sensors, it is important to understand their working principle, as well as their strength and weakness. In this chapter, we provide comprehensive information on how depth sensors track human facial and body features using computer vision and pattern recognition based techniques, and identify their strength in computational cost and robustness. Then, we focus on the major weakness of depth sensors, that is, the low accuracy that happens during occlusion, and explain possible solutions to improve recognition quality in detail. In particular, we discuss in depth on machine learning-based reconstruction method that utilize prior knowledge to correct corrupted data obtained by the sensors. Finally, we give some examples on depth sensors-based application, especially in the field of animation creation, to show how these sensors can improve existing methods in human-computer interaction.

In the rest of this chapter, we review the state of the art in section "State of the Art." We explain in more details about how depth sensors obtain and process human facial and body movement information in section "Extracting Facial and Body Information." We then discuss on the main challenge on depth sensor-based systems, which is about the relatively low accuracy of the data obtained, and explain how this challenge can be tackled in section "Dealing with Noisy Data." We finally point out various emerging applications developed with depth sensors in section "Depth Camera-Based Applications" and conclude this chapter in section "Conclusion."

State of the Art

Typical depth sensors utilize a depth camera to obtain a depth image. The main advantage of the depth camera over traditional color cameras is that instead of obtaining color information, it estimates the distance of the objects seen by the

camera using an infrared sensor. The images taken from a depth camera are called depth images. In these images, the pixels represent distance instead of color.

The nature of depth images provides a huge advantage in automatic recognition using computer vision and machine learning algorithm. With traditional color images, recognizing objects requires segmenting them based on color information. This is challenging under situations in which the background has a similar color as the foreground objects (Fernandez-Sanchez et al. 2013). Moreover, color values are easily affected by lighting conditions, which reduces the robustness of object recognition (Kakumanu et al. 2007). On the contrary, with depth images, since the pixel value represents distance, automatic object segmentation becomes independent of the color of the object. As long as the object is geometrically separated from the background, accurate segmentation can be performed. Followed by such an improved segmentation process is an improved object recognition system, which identifies the nature of the objects using accurate geometric features. Such an advancement in accuracy and robustness allows depth sensors to become a popular commercial product that leads to many new applications.

The Microsoft Kinect (https://developer.microsoft.com/en-us/windows/kinect), which utilizes both color and depth cameras, is one of the most popular depth sensors. Due to the uses of both color and depth cameras, Kinect can create a 3D point cloud based on the obtained images. Figure 1 shows the images obtained by the two cameras, as well as two views of the corresponding point cloud. Kinect gaming usually involves players controlling the gameplay with body movement. Virtual characters in the game are then synthesized on the fly based on the movement information obtained. Such a kind of application involves different domains of research. First, computer vision and machine learning techniques are applied to analyze the depth images obtained by the depth sensor. This typically involves recognizing different human features, such as the human body parts (Shotton et al. 2012). Then, human computer interaction researches are applied to translate to control signals from the body movement into gameplay controls. Computer graphics and animation algorithms are used to create real-time rendering, which usually includes character animation synthesized from the movement of the player. In some situations, virtual reality (Kyan et al. 2015) or augmented reality (Vera et al. 2011) research is adapted to enhance the immersiveness of the game.

However, depth sensors are not without their weaknesses. Comparing to traditional capturing devices such as accelerometers, the accuracy of depth sensors is

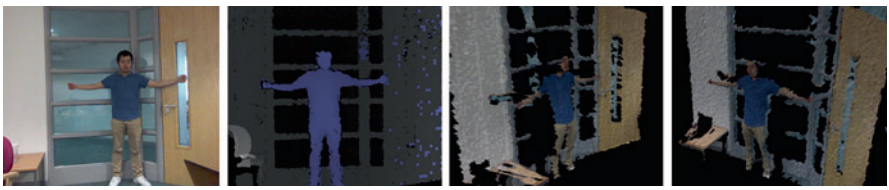

Fig. 1 (From *left* to *right*) The color and depth images obtained by a Microsoft Kinect, as well as two views of 3D point cloud rendered by combing the color and depth information

considerably lower. This is mainly because these sensors usually consist of a single depth camera. When occlusions occur, the sensors cannot obtain information from the shielded area. This results in a significant drop in recognition accuracy. While it is possible to utilize multiple depth cameras to obtain better results, one has to deal with the cross-talk, interference of infrared signals, among multiple cameras (Alex Butler et al. 2012). It also deficits the advantage of using depth sensors in terms of easy setup and efficient capture. Therefore, it is preferable to enhance the sensor accuracy using software algorithms, instead of introducing more hardware.

To enhance the quality of the obtained data, machine learning approaches using prior knowledge of the face and body features have shown great success (Shum et al. 2013). The main idea is to apply prior knowledge onto the tracked data and correct the less reliable parts or introduce more details onto the data. Such knowledge can either be defined manually or learned from examples. The key here is to represent the prior knowledge in a way that is efficient and effective to be used during run-time.

Extracting Facial and Body Information

There is a large body of research in obtaining facial and body information from depth cameras. In this section, we explain some of the main methods and discuss on their performances.

Facial Feature

Facial feature detection usually involves face segmentation and landmark detection. The former segments the face from the background, while the latter detect key regions and feature points. To segment the face area from the background and the rest of the human body, one can detect the skin color and perform segmentation (Bronstein et al. 2005). However, such a method is easily affected by illumination. Using the histogram of depth information from the depth image can improve the system robustness (Segundo et al. 2010). Since human faces have the same topology, it is possible to apply geometric rules to identify landmarks on the face. A simple example is to approximate the face with an ellipse and divide the ellipse into different slices based on predefined angles (Segundo et al. 2010). For each slices, corresponding features can be searched based on the 3D height of the face. For example, the eyes are the lowest point on the corresponding slice, while the nose is the highest. Similarly, it is possible to use local curvature to represent different features on the face, so as to determine different facial regions (Chang et al. 2006). For example, the eye regions are usually a valley and can be represented by a specific value of mean curvature and Gaussian curvature. The disadvantage for these methods is that manually defined geometric rules may not be robust for different users, especially for users coming from different countries. A better solution is to apply a data-driven approach. For example, one can construct a database with

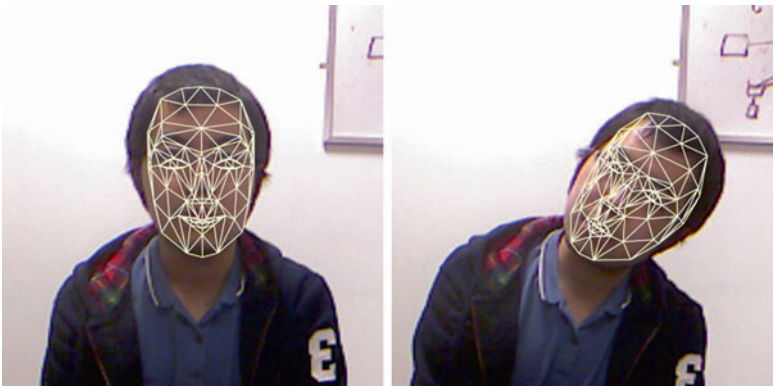

Fig. 2 3D facial landmark identified by Kinect overlapped on 2D color images

segmented facial regions and train a random forest that can automatically identify facial regions on a face (Kazemi et al. 2014).

Another direction of representing facial features is to use predefined facial template (Li et al. 2013; Weise et al. 2011). Such a template is a high-quality 3D mesh with controllable parameters. During run-time, the system deforms the 3D template to align with the geometry structure of the segmented face from the depth image. Such a deformation process is usually done by numerical optimization due to the high degree of freedom. Upon successful alignment, the systems can understand the observed face in the depth image with the deformed template. It can also represent the face with a set of deformation parameters so as to control animation in real-time.

Microsoft Kinect also provides support for 3D landmark detection, as shown in Fig. 2. Different expressions can be identified based on the arrangements of 3D landmarks. Such understanding of facial orientation and expression is useful for real-time animation control.

Body Posture

The mainstream of depth sensors-based body recognition system is to apply pattern recognition and machine learning techniques to identify the human subject. By training a classifier that can identify how individual body part appears in the depth image, one can recognize these parts using real-time depth camera input (Girshick et al. 2011; Shotton et al. 2012; Sun et al. 2012). There are several major challenges in this algorithm. Chief among them is the availability of training data. In order to train a classifier, a large number of depth images with annotation indicating the body parts are needed. Since body parts appear differently based on the viewing angle, the training database should capture such parts in different viewpoints. Moreover, since users of different body sizes appear differently in depth images, to train a robust classifier that can handle all users, training images consisting of body variation are

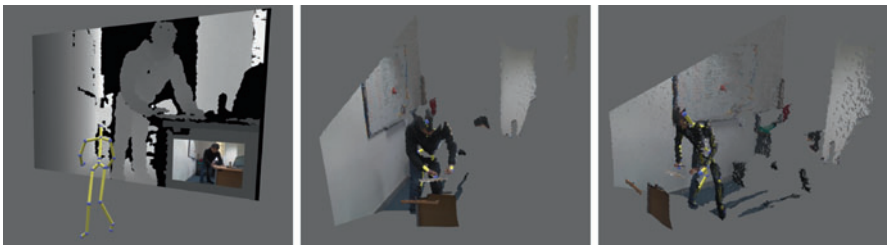

Fig. 3 (*Left*) The 3D skeleton obtained by Microsoft Kinect with the corresponding depth and color images. (*Middle* and *Right*) Two views of the 3D point cloud together with the obtained 3D skeleton

needed. As a result, hundreds of thousands of annotated depth images will be required, which exceeds what human labors can generate. To solve the problem, it is proposed to synthesize depth images using different humanoid models and 3D motion capture data. Since the body part information of these humanoid models is known in advance, it becomes possible to automatically annotate the position of the body parts in the synthesized depth image. With these training images, one can train a decision forest to classify depth pixels into the corresponding body parts. Different designs of decision forest have resulted in different level of success, and they are all capable of identifying body parts in real-time.

Microsoft Kinect also applied a pattern recognition approach to recognize body parts with the depth images (Shotton et al. 2012). Figure 3 shows the results of Kinect posture recognition, which is shown as the yellow skeleton. By overlapping the skeleton with the 3D point cloud, it can be observed that the Kinect performs reasonably accurate under normal circumstances.

Another stream of method in body identification and modeling is to take advantage of the geometry of human body and utilize body template model (Liu et al. 2016a; Zhang et al. 2014a). First, the pixels in the depth image that belongs to the human body are extracted. Since the pixel value represents distance, one can project them into a 3D space and create a point cloud of human body. Then, the system fits a 3D humanoid mesh model into such a point cloud, so as to estimate the body posture. This process involves deforming the 3D mesh model such that the surface of the model aligns with the point cloud. Since the template model contains human information such as body parts, when deforming the model to fit the point cloud, we identify the corresponding body information in the point cloud. The main challenge in this method is to deform the mesh properly to avoid unrealistic postures and over-deformed surfaces, which is still a challenging research problem. Physics-based motion optimization can ensure the physical correctness of the generated postures (Zhang et al. 2014a). Utilizing simplified, intermediate template for deformation optimization can enhance the optimization performance (Liu et al. 2016a). This method can potentially provide richer body information depending on the template used. However, a major drawback of such an optimization-based approach is the higher run-time computational cost, making it inefficient to be applied in real-time systems.

Human Environment Interaction

The depth images captured do not only contain information about the user but also the surrounding environment. Therefore, it is possible to identify high-level information about how the user interacts with the environment.

Unlike the human body, the environment does not have a uniform structure, and therefore it is not possible to fit a predefined template or apply prior knowledge. Geometry information based on planes and shapes become the next available information to extract. The RANdom SAmple Consensus (RANSAC) algorithm can be used to identify planer objects in the scenes such as walls and floors, which can help to understand how the human user moves around in the open areas (Mackay et al. 2012). It is also possible to compare successive depth images to identify the moving parts, in order to understand how the user interacts with external objects (Shum 2013).

Depth cameras can be used for 3D scanning in order to obtain surface information of the environment or even the human user. While one depth image only provide information about a partial surface, which we called a 2.5D point cloud, multiple depth images taken from different viewing angle can combine and form a full 3D surface. One of the most representative systems in this area is called KinectFusion (Newcombe et al. 2011). Such a system requires the user to carry a Kinect and capture depth images continuously over a static environment. Real-time registration is performed to understand the 3D translation and rotation movement of the depth camera. This allows alignment of multiple depth images to form a complete 3D surface. Apart from scanning the environment, it is possible to scan the face and body of a human user (Cui et al. 2013) and apply real-time posture deformation on the Kinect tracked skeleton (Iwamoto et al. 2015). Finally, because single view depth cameras suffer from the occlusion problem, it is proposed to capture how human users interact with objects by combining KinectFusion, color cameras, and accelerometer-based motion capture system (Sandilands et al. 2012, 2013).

Since depth sensors can obtain both environment and human information, it facilitates the argument that human information can enhance understanding of unstructured environment (Jiang and Saxena 2013; Jiang et al. 2013). Using a chair as an example. A chair can come with different shapes and designs, which makes recognition extremely difficult. However, the general purpose of a chair is for human to rest on. Therefore, with the human movement information obtained by depth cameras, we can identify a chair not just by its shapes but also by the way the human interacts with it. Similarly, human movement may be ambiguous sometimes. Understanding the environment helps us to identify the correct meaning of the human motion. Depth sensors open up new directions on recognition by considering human and environment information together.

Dealing with Noisy Data

The main problem of using depth sensors is to deal with the noisy data obtained. In particular, most depth sensor-based applications rely on a single point of view to obtain the depth image. As a result, the quality of the detected face and posture are of

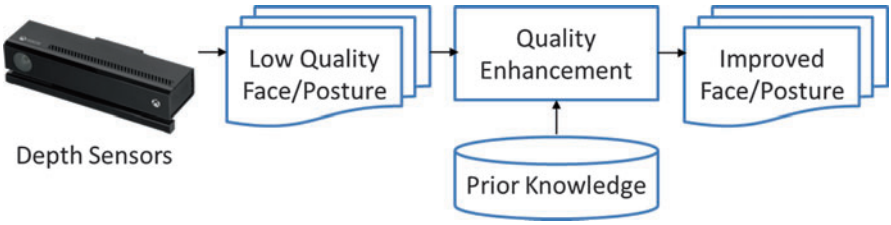

Fig. 4 An overview of depth sensors data enhancement

low resolution and suffer heavily from occlusion. It is possible to apply machine learning algorithms to enhance the quality of the data. The idea is to introduce a quality enhancement process that considers prior knowledge of the human body, which is typically a database of high-quality faces or postures, as shown in Fig. 4. In this section, we discuss how body and facial information can be reconstructed from noisy data.

Face Enhancement

While depth sensors can obtain facial features, due to the relatively low resolution, the quality of the features is not always satisfying. The 3D face obtained is usually missing details and unrealistic. In this section, we explain how we can enhance the quality of 3D faces obtained from depth sensors.

Since the quality of a single depth image is usually noisy with low resolution, the 3D facial surface generated is rough. By obtaining high-quality 3D faces through 3D scanners and their corresponding color texture, one can construct a face database and extract the corresponding prior knowledge (Liang et al. 2014; Wang et al. 2014). These faces in the database are divided into patches such as eyes, nose, etc. Since color texture is available, one can take advantage of color features to enhance the segmentation accuracy. Given the low quality depth and color images of a face obtained from sensors, facial regions are obtained in run-time. For each region obtained, a set of similar patches is found in the database. Such a region is then approximated by a weight sum of the database patches. By replacing different parts of the run-time face image with their corresponding approximation, a high-quality 3D face surface can be generated. This method depends heavily on the quality and variety of face in the database, as well as the way we abstract those faces to represent the one observed by depth sensors in run-time.

Constructing a database for prior knowledge is costly. It is therefore proposed to scan the face of the user in different angles, and apply such a face to enhance the run-time detected face (Zollhöfer et al. 2014). The system first requests the user to rotate around a depth sensor and obtain a higher quality 3D mesh, using registration methods similar to the KinectFusion mentioned in the last section (Newcombe et al. 2011). Then, given a run-time lower quality depth image of the face, the system deforms the high quality 3D mesh such that it aligns with the depth image

pixels. As a result, high quality mesh with run-time facial expression can be generated. The core problem here is to deform the high quality facial mesh nicely and avoid generating visual artifact. It is shown that by dividing the face into multiple facial regions to strengthen the feature correspondence, deformation quality can be improved (Kazemi et al. 2014).

Posture Enhancement

The body tracked by depth sensors may contain inaccurate body parts due to different types of error. Simple sensor error can be caused by geometry shape of body parts and viewing angles. It is proposed to apply Butterworth filter (Bailey and Bodenheimer 2012) or a simple low-pass filter (Fern'ndez-Baena et al. 2012) to smooth out the vibration effect of tracked positions due to this type of error. However, when occlusions occur, in which a particular body part is shield from the camera, the tracked body position would contain a large amount of error. Simple filter will not be sufficient to correct these postures.

As a solution, it is proposed to utilize accurately captured 3D human motion as prior knowledge and reconstruct the inaccurate postures from the depth sensor. In this method, a motion database is constructed using carefully captured 3D motion, usually with optical motion capture systems. Given a depth sensor posture, one can search for a similar posture in the database. The missing or error body parts from the depth sensors can be replaced by those in the corresponding database posture (Shum and Ho 2012). However, such a naive method cannot perform well for complex posture, as using only one posture from the database cannot always generalize the posture performed by the user, and therefore cannot effectively reconstruct the posture.

More advanced posture reconstruction algorithms utilize machine learning to generalize posture information from the motion database (Chai and Hodgins 2005; Liu et al. 2011; Tautges et al. 2011). In particular, the motion database is used to create a low dimensional latent space by dimensionality reduction techniques. Since the low dimensional space is generated using data from real human, each point in the space represents a valid natural posture. Given a partially mistracked posture from a depth camera, one can project the posture into the learned low dimensional space and apply numerical optimization to enhance the quality of the posture. The optimized result is finally back-projected into a full body posture. Since the optimization is performed in the low dimensional latent space, the solution found should also be a natural posture. In other words, the unnatural elements due to sensor error can be removed. The major problem of this method is that the system has no information about which part of the body posture is incorrect. Therefore, while one would expect the system to correct the error parts of the posture using information from the accurate parts, the actual system may perform vice versa. As a result, the optimized posture may no longer be similar to the original depth sensor input.

To solve the problem, optimization process that considers the reliability of individual body part is proposed (Shum et al. 2013). The major difference from

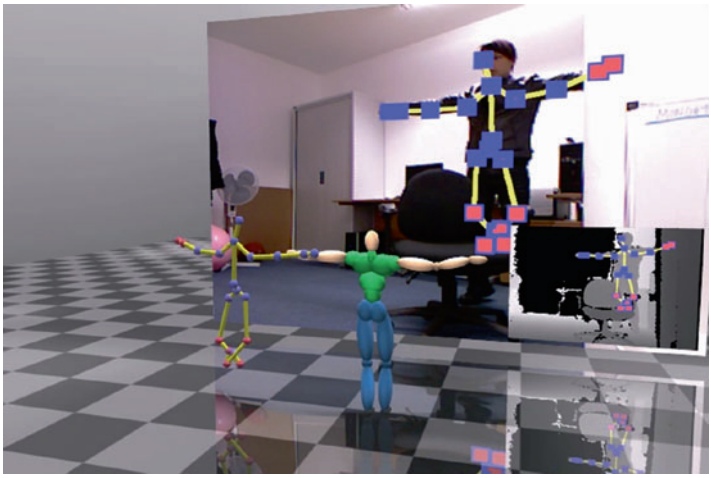

Fig. 5 Applying posture reconstruction to enhance the quality of the obtained data

this method comparing with prior ones is that it divide the posture reconstruction process into two steps. In the first step, a procedural algorithm is used to evaluate the degree of reliability of individual body parts. This is by accessing the behavior of a tracked body part to see if the position of the part is inconsistent, as well as accessing the part with respect to its neighbor body parts to see if it creates inconsistent bone length. In the second step, posture reconstruction is performed with reference to this reliability information, such that the system relies on the more parts with higher reliability. Essentially, the reliability information helps the system to explicitly use the correct body parts and reconstruct the incorrect ones. Such a system can be further improved by using Gaussian process to model the motion database, which helps to reduce the amount of motion data needed to reconstruct the posture (Liu et al. 2016b; Zhou et al. 2014). Better rules to estimate the reliability of the body parts can also enhance the system performance (Ho et al. 2016).

Figure 5 shows the result of applying posture reconstruction. The color and depth images show that the user is occluded by a chair and the surrounding environment. The yellow skeleton on the left is the raw posture obtained by Kinect, in which less reliable body parts are highlighted in red. The right character shows the reconstructed posture using the method proposed in (Shum et al. 2013). The awkward body parts are identified and corrected using the knowledge learned from a motion database.

Prior Knowledge

The major research focus of face and posture enhancement is to apply appropriate prior knowledge to improve data obtained in rum-time. In machine learning-based

algorithms, such prior knowledge is usually learned from a database and represented in a format that can be efficiently used in run-time.

For motion enhancement, since human-motion is highly nonlinear with large variation, it is not effective to represent the database using a single model. Instead, many of the existing research apply multiple local models to represent the database, such as using a mixture of Gaussian model (Liu et al. 2016b). It is also proposed to apply deep learning to learn a set of manifolds that represents a motion database (Holden et al. 2015). Precomputing these models and manifolds are time-consuming, as it involves abstracting the whole database. Therefore, lazy learning algorithm is adapted, in which modeling of the database is not done as a preprocesss but as a run-time process using run-time information (Chai and Hodgins 2005; Shum et al. 2013). During run-time, based on the user-performed posture, the system retrieves a number of relevant postures from the database and models such a subset of postures only. This method has two advantages. First, by modeling only a small number of postures that are relevant to the performed posture, one can reduce the computational cost of constructing a latent space. Second, since the subset of postures are relatively similar, one can assume that they all lay in a locally linear space and apply simpler linear dimensionality reduction to generate the latent space. This allows real-time generation of the latent space. With improved database organization, the database search time can be further reduced and the relevancy of the retrieved results can be enhanced (Plantard et al. 2016a, b), such that real-time ergonomic and motion analysis applications can be preformed (Plantard et al. 2016b).

Figure 6 visualizes how prior knowledge can be estimated from database. Each blue circle in the figure represents a database entry, and the filling color represents its value. The obtained prior knowledge from the scattered database entries is represented by the shaded area, which enables one to understand the change of value within the considered space. The left figure shows a traditional machine learning algorithm, in which prior knowledge is obtained as a preprocess, considering all database entries. During run-time, when a query arrives, the system uses the knowledge to estimate the corresponding value of the query. The right figure shows the case of lazy learning, in which prior knowledge is obtained during run-time. This allows the system to extract database entries that are more similar to the query and estimate the prior knowledge with only such a subset of data.

Depth Camera-Based Applications

With depth sensors, it becomes possible to consider the user posture as part of an animation system and create real-time animation. Here, we discuss on some depth sensors-based animation control systems and point out the challenges and solutions.

Producing real-time facial animation with depth sensors is efficient. By representing the facial features with a deformed template, it is possible to drive the facial expression of virtual 3D faces (Li et al. 2013; Weise et al. 2011). Due to the different dimensions between the faces of the user and the character, directly

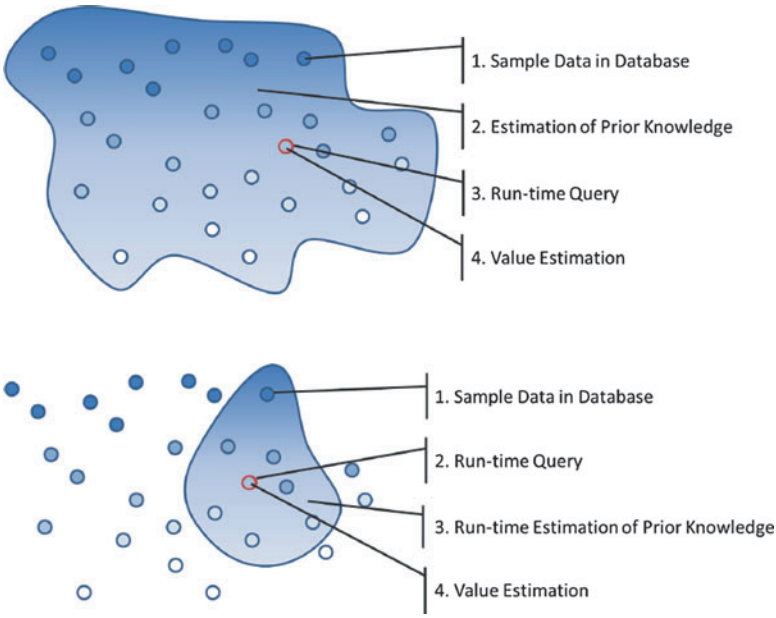

Fig. 6 (*Upper*) Traditional machine learning that represents the prior from the whole database. (*Lower*) Lazy learning that represents the prior from a subset of the database based on the online query

applying facial features such as landmark locations generates suboptimal results. The proposed common template acts as a bridge to connect the two ends. Such a template is a parametric representation of the face, which is more robust against difference in dimensions. With the template, it becomes possible to retarget the user's expression onto the character's face.

Typical real-time animation systems such as games utilize a motion database to understand what the user performs and renders the scenario accordingly. For example, one can compare the user performed motion obtained from depth sensors with a set of motion in the database and understand the nature of the motion as well as how it should affect the real-time render (Bleiweiss et al. 2010). Alternatively, with an interaction database, one can generate a virtual character that acts according to the posture of the user, in order to create a two character dancing animation, which is difficult to be captured due to hardware limitation (Ho et al. 2013).

While it is possible to utilize the posture captured by depth sensors for driving the animation of virtual characters, the generated animation may not be physically correct and dynamically plausible. On the one hand, since the depth sensors track kinematic positions only, there is no information about the forces exerted. It is proposed to combine the uses of depth cameras with pressure sensors and estimate the internal joint torque using inverse dynamics (Zhang et al. 2014b). This allows simulating virtual characters with physically correct movement. On the other hand, while depth sensors can track the body parts positions, it is relatively difficult to track

Fig. 7 Real-time dynamic deformed character generated from Kinect postures

how the body deforms dynamically during the movement. Therefore, it is proposed to enhance the realism of the generated character by applying real-time physical simulation onto Kinect postures (Iwamoto et al. 2015). This allows the system to synthesize real-time dynamic deformation, such as the jiggling of fresh, based on the movement obtained in real-time, as shown in Fig. 7.

Utilizing depth cameras, user can interact with virtual objects with body motion. On the one hand, predefined hand and arm gestures can be used to control virtual objects. Once the Kinect has detected a set of specific gestures, a 3D virtual object can be fitting onto the user and move according to the user's gesture (Soh et al. 2013). On the other hand, the virtual objects can be attached on the user's body and move with the user's posture, such as carrying a virtual handbag (Wang et al. 2012).

Depth sensors fuse a new application known as the virtual fitting, in which the shopping experience can be facilitated by letting customers to try on virtual clothing. This allows mix-and-match of clothes and accessories in real-time without being physically present in the retail shop. The system involves building a 2D segmented clothing database indexed by the postures of the user. During run-time, the system searches for suitable database entries and overlays them on the customers fitting image (Zhou et al. 2012). Another clothes fitting method is to utilize a 3D clothing database and obtain 3D models of the user with depth sensors. This allows the system to recommend items that fits with the user's body (Pachoulakis and Kapetanakis 2012).

Conclusion

In this chapter, we explained how depth sensors are applied to gather human facial and body posture information in generating and controlling animations. Depth sensors obtain human information in real-time and provide a cheaper alternative for human motion capturing. However, there are still rooms to improve the quality of the obtained data. In particular, depth sensors suffer heavily from occlusions, in which part of the human body is shield. Machine learning algorithms can reconstruct the data and improve the quality, but more research is needed to solve the problem. Depth sensors fuse many interesting applications in the computer animation and

games domains, providing real-time control on animation creation. Given the rate of new depth sensors research and applications, such a technology can become an important part of the daily life in the near future.

Acknowledgment This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) (Ref: EP/M002632/1).

References

Alex Butler D, Izadi S, Hilliges O, Molyneaux D, Hodges S, Kim D (2012) Shake'n'sense: reducing interference for overlapping structured light depth cameras. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI'12. ACM, New York, pp 1933–1936

Bailey SW, Bodenheimer B (2012) A comparison of motion capture data recorded from a vicon system and a Microsoft Kinect sensor. In: Proceedings of the ACM symposium on applied perception, SAP'12. ACM, New York, pp 121–121

Bleiweiss A, Eshar D, Kutliroff G, Lerner A, Oshrat Y, Yanai Y (2010) Enhanced interactive gaming by blending full-body tracking and gesture animation. In: ACM SIGGRAPH ASIA 2010 Sketches. Seoul, South Korea. ACM, p 34

Bronstein AM, Bronstein MM, Kimmel R (2005) Three-dimensional face recognition. Int J Comput Vision 64(1):5–30

Chai J, Hodgins JK (2005) Performance animation from low-dimensional control signals. In SIGGRAPH'05: ACM SIGGRAPH 2005 Papers. ACM, New York, pp 686–696

Chang KI, Bowyer KW, Flynn PJ (2006) Multiple nose region matching for 3d face recognition under varying facial expression. IEEE Trans Pattern Anal Mach Intell 28(10):1695–700

Cui Y, Chang W, Nöll T, Stricker D (2013) Kinectavatar: fully automatic body capture using a single Kinect. In: Proceedings of the 11th international conference on computer vision, vol 2, ACCV'12. Springer-Verlag, Berlin/Heidelberg, pp 133–147

Fern'ndez-Baena A, SusÃn A, Lligadas X (2012) Biomechanical validation of upper-body and lower-body joint movements of Kinect motion capture data for rehabilitation treatments. In: Intelligent Networking and Collaborative Systems (INCoS), 2012 4th International Conference on, pp 656–661

Fernandez-Sanchez EJ, Diaz J, Ros E (2013) Background subtraction based on color and depth using active sensors. Sensors 13(7):8895–915

Girshick R, Shotton J, Kohli P, Criminisi A, Fitzgibbon A (2011) Efficient regression of general-activity human poses from depth images. In: Computer Vision (ICCV), 2011 I.E. international conference on. Barcelona, Spain. pp 415–422

Ho ESL, Chan JCP, Komura T, Leung H (2013) Interactive partner control in close interactions for real-time applications. ACM Trans Multimedia Comput Commun Appl 9(3):21:1–21:19

Ho ES, Chan JC, Chan DC, Shum HP, Cheung YM, Yuen PC (2016) Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. Comput Vis Image Underst 148:97–110. doi:10.1111/cgf.12735

Holden D, Saito J, Komura T, Joyce T (2015) Learning motion manifolds with convolutional autoencoders. In ACM SIGGRAPH ASIA 2015 technical briefs. ACM, Kobe, Japan. 2015 SIGGRAPH ASIA

Iwamoto N, Shum HPH, Yang L, Morishima S (2015) Multi-layer lattice model for real-time dynamic character animation. Comput Graph Forum 34(7):99–109

Jiang Y, Saxena A (2013) Hallucinating humans for learning robotic placement of objects. In: Proceedings of the 13th international symposium on experimental robotics. Springer International Publishing, Heidelberg, pp 921–937

Jiang Y, Koppula H, Saxena A (2013) Hallucinated humans as the hidden context for labeling 3d scenes. In: Proceedings of the 2013 I.E. conference on computer vision and pattern recognition, CVPR'13. IEEE Computer Society, Washington, DC, pp 2993–3000

Kakumanu P, Makrogiannis S, Bourbakis N (2007) A survey of skin-color modeling and detection methods. Pattern Recogn 40(3):1106–22

Kazemi V, Keskin C, Taylor J, Kohli P, Izadi S (2014) Real-time face reconstruction from a single depth image. In: 3D Vision (3DV), 2014 2nd international conference on, vol 1. IEEE, Lyon, France. 2014 3DV. pp 369–376

Kinect sdk. https://developer.microsoft.com/en-us/windows/kinect

Kyan M, Sun G, Li H, Zhong L, Muneesawang P, Dong N, Elder B, Guan L (2015) An approach to ballet dance training through ms Kinect and visualization in a cave virtual reality environment. ACM Trans Intell Syst Technol (TIST) 6(2):23

Li H, Yu J, Ye Y, Bregler C (2013) Realtime facial animation with on-the-fly correctives. ACM Trans Graph 32(4):42–1

Liang S, Kemelmacher-Shlizerman I, Shapiro LG (2014) 3d face hallucination from a single depth frame. In: 3D Vision (3DV), 2014 2nd international conference on, vol 1. IEEE, Lyon, France. 2014 3DV. pp 31–38

Liu H, Wei X, Chai J, Ha I, Rhee T (2011) Realtime human motion control with a small number of inertial sensors. In: Symposium on interactive 3D graphics and games, I3D'11. ACM, New York, pp 133–140

Liu Z, Huang J, Bu S, Han J, Tang X, Li X (2016a) Template deformation-based 3-d reconstruction of full human body scans from low-cost depth cameras. IEEE Trans Cybern PP(99):1–14

Liu Z, Zhou L, Leung H, Shum HPH (2016b) Kinect posture reconstruction based on a local mixture of gaussian process models. IEEE Trans Vis Comput Graph 14 pp. doi:10.1109/TVCG.2015.2510000

Mackay K, Shum HPH, Komura T (2012) Environment capturing with Microsoft Kinect. In: Proceedings of the 2012 international conference on software knowledge information management and applications, SKIMA'12. Chengdu, China. 2012 SKIMA

Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) Kinectfusion: real-time dense surface mapping and tracking. In: Proceedings of the 2011 10th IEEE international symposium on mixed and augmented reality, ISMAR'11. IEEE Computer Society, Washington, DC, pp 127–136

Pachoulakis I, Kapetanakis K (2012) Augmented reality platforms for virtual fitting rooms. Int J Multimedia Appl 4(4):35

Plantard P, Shum HP, Multon F (2016a) Filtered pose graph for efficient kinect pose reconstruction. Multimed Tools Appl 1–22. doi:10.1007/s11042-016-3546-4

Plantard P, Shum HPH, Multon F (2016b) Ergonomics measurements using Kinect with a pose correction framework. In: Proceedings of the 2016 international digital human modeling symposium, DHM '16, Montreal, 8 p

Sandilands P, Choi MG, Komura T (2012) Capturing close interactions with objects using a magnetic motion capture system and a rgbd sensor. In: Proceedings of the 2012 motion in games. Springer, Berlin/Heidelberg, pp 220–231

Sandilands P, Choi MG, Komura T (2013) Interaction capture using magnetic sensors. Comput Anim Virtual Worlds 24(6):527–38

Segundo MP, Silva L, Bellon ORP, Queirolo CC (2010) Automatic face segmentation and facial landmark detection in range images. Systems Man Cybern Part B Cybern IEEE Trans 40 (5):1319–30

Shotton J, Girshick R, Fitzgibbon A, Sharp T, Cook M, Finocchio M, . . . Blake A (2013) Efficient human pose estimation from single depth images. IEEE Trans Pattern Anal Machine Intell 35 (12):2821–2840

Shum HPH (2013) Serious games with human-object interactions using rgb-d camera. In: Proceedings of the 6th international conference on motion in games, MIG'13. Springer-Verlag, Berlin/Heidelberg

Shum HPH, Ho ESL (2012) Real-time physical modelling of character movements with Microsoft Kinect. In: Proceedings of the 18th ACM symposium on virtual reality software and technology, VRST'12. ACM, New York, pp 17–24

Shum HPH, Ho ESL, Jiang Y, Takagi S (2013) Real-time posture reconstruction for Microsoft Kinect. IEEE Trans Cybern 43(5):1357–69

Soh J, Choi Y, Park Y, Yang HS (2013) User-friendly 3d object manipulation gesture using Kinect. In: Proceedings of the 12th ACM SIGGRAPH international conference on virtual-reality continuum and its applications in industry, VRCAI'13. ACM, New York, pp 231–234

Sun M, Kohli P, Shotton J (2012) Conditional regression forests for human pose estimation. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. conference on. Providence, Rhode Island. pp 3394–3401

Tautges J, Zinke A, Krüger B, Baumann J, Weber A, Helten T, Müller M, Seidel H-P, Eberhardt B (2011) Motion reconstruction using sparse accelerometer data. ACM Trans Graph 30 (3):18:1–18:12

Vera L, Gimeno J, Coma I, Fernández M (2011) Augmented mirror: interactive augmented reality system based on Kinect. In: Human-Computer Interaction–INTERACT 2011. Springer, Lisbon, Portugal. 2011 INTERACT. pp 483–486

Wang L, Villamil R, Samarasekera S, Kumar R (2012) Magic mirror: a virtual handbag shopping system. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 I.E. computer society conference on. IEEE, Rhode Island. 2012 CVPR. pp 19–24

Wang K, Wang X, Pan Z, Liu K (2014) A two-stage framework for 3d facereconstruction from rgbd images. Pattern Anal Mach Intell IEEE Trans 36(8):1493–504

Weise T, Bouaziz S, Li H, Pauly M (2011) Realtime performance-based facial animation. ACM Trans Graph (TOG) 30:77, ACM

Zhang P, Siu K, Jianjie Z, Liu CK, Chai J (2014a) Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. ACM Trans Graph 33(6):221:1–221:14

Zhang P, Siu K, Jianjie Z, Liu CK, Chai J (2014b) Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. ACM Trans Graph (TOG) 33(6):221

Zhou Z, Shu B, Zhuo S, Deng X, Tan P, Lin S (2012) Image-based clothes animation for virtual fitting. In: SIGGRAPH Asia 2012 technical briefs. ACM, Singapore. 2012 SIGGRAPH ASIA. p 33

Zhou L, Liu Z, Leung H, Shum HPH (2014) Posture reconstruction using Kinect with a probabilistic model. In: Proceedings of the 20th ACM symposium on virtual reality software and technology, VRST'14. ACM, New York, pp 117–125

Zollhöfer M, Nießner M, Izadi S, Rehmann C, Zach C, Fisher M, Wu C, Fitzgibbon A, Loop C, Theobalt C et al (2014) Real-time non-rigid reconstruction using an rgb-d camera. ACM Trans Graph (TOG) 33(4):156