sensors

Article

A Timestamp-Independent Haptic–Visual Synchronization Method for Haptic-Based Interaction System

Yiwen Xu [1,2], Liangtao Huang [1], Tiesong Zhao [1], Ying Fang [1] and Liqun Lin [1,*]

[1] Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China; xu_yiwen@fzu.edu.cn (Y.X.); 211120101@fzu.edu.cn (L.H.); t.zhao@fzu.edu.cn (T.Z.); fangying@fzu.edu.cn (Y.F.)

[2] College of Zhicheng, Fuzhou University, Fuzhou 350108, China

[*] Correspondence: lin_liqun@fzu.edu.cn

Abstract: The booming haptic data significantly improve the users' immersion during multimedia interaction. As a result, the study of a Haptic-based Interaction System has attracted the attention of the multimedia community. To construct such a system, a challenging task is the synchronization of multiple sensorial signals that is critical to the user experience. Despite audio-visual synchronization efforts, there is still a lack of a haptic-aware multimedia synchronization model. In this work, we propose a timestamp-independent synchronization for haptic–visual signal transmission. First, we exploit the sequential correlations during delivery and playback of a haptic–visual communication system. Second, we develop a key sample extraction of haptic signals based on the force feedback characteristics and a key frame extraction of visual signals based on deep-object detection. Third, we combine the key samples and frames to synchronize the corresponding haptic–visual signals. Without timestamps in the signal flow, the proposed method is still effective and more robust in complicated network conditions. Subjective evaluation also shows a significant improvement of user experience with the proposed method.

Keywords: haptic-based interaction system; multimedia environment; human-centric multimedia; haptic–visual synchronization

check for updates

Citation: Xu , Y.; Huang, L.; Zhao, T.; Fang, Y.; Lin, L. A Timestamp-Independent Haptic–Visual Synchronization Method for Haptic-Based Interaction System. *Sensors* **2022**, *22*, 5502. https://doi.org/10.3390/s22155502

Academic Editor: Stefano Berretti

Received: 17 June 2022
Accepted: 21 July 2022
Published: 23 July 2022

1. Introduction

Recent developments in multimedia technology also require multimedia content that is more immersive. As an emerging multimedia signal, haptics provide newfangled and authentic user experiences beyond current audio-visual signals. Thus, a Haptic-based Interaction System (HIS) has garnered the attention of researchers [1–5].

An HIS has been used in a variety of applications. For example, Ilaria et al. [6] designed an immersive haptic VR system for rehabilitation training of children with motor neurological disorders which significantly improved the effect of rehabilitation training. Zhou et al. [7] proposed an approach with visual and haptic signals which helps physicians perform surgeries accurately and effectively and furthermore reduces their physical and cognitive burden during surgery. Chen et al. [8] designed a remote training system with force feedback for power grid operation training. It avoided the collision between the manipulator and steel bars, which helps guide operators reduce operational errors and complete tasks efficiently. Varun et al. [9] also introduced haptics into a VR-based training system to enhance training immersion, effectiveness and efficiency. The use ofan HIfor online shopping [10,11] can improve the realism of the shopping experience and help visually impaired patients enjoy the convenience of online shopping. An HIS can also be used in outdoor search and rescue scenarios to avoid collisions by providing tactile guidance [12]. In industry, an HIis usually used to enhance the operational ability of robots. For example, the work in [13] equipped a robot with bionic haptic manipulators to help it have more stable grasping ability in tele-operation tasks. In [14], the operator controls the robot to perform the tele-operation in real time by means of a pneumatic haptic feedback glove. Apparently, the HIS is widely used and worthy of further investigation.

In an HIS, similar to conventional audio-visual signals, the haptic signal can also be affected during network fluctuations or congestion. In a multimedia case, the haptic signal may lose synchronization with other signals, e.g., images and videos. Compared to video, audio or image, the transmission of haptics is more tolerant of data loss and bandwidth but has higher requirements for the latency between signals. To ensure more natural interactive operations, haptic-based multimedia signal transmission requires better inter-signal synchronization. As reported, the haptic–visual asynchronization greatly influences the user experience. Qi et al. [15] implemented several experiments to explore the impact of the delay between video and haptic signals on the quality of users' experience. The results showed that all the Mean Opinion Score (MOS) values decreased with the inter-flow synchronization error. The works from Aung et al. [16] also confirmed the above conclusion.

To address this issue, haptic–visual synchronization is needed. The system examines the synchronization status of signals in real time and adjusts the corresponding signals immediately when an asynchronization is found. However, to the best of our knowledge, current research on the synchronization of visual–haptic signals is mainly focused on studying the impact of visual–haptic asynchronization on user experience, while little research has been conducted on synchronization detection and adjustment of visual–haptic signals, and there is still room for improvement in this area.

The research on synchronization algorithms for audio-visual signals can be used as good references for the research on visual–haptic signals. In the state-of-the-art HIS systems, synchronization is achieved by the timestamp method [17,18] that was designed for generic signals. The timestamp-dependent method embeds the timestamps in the signal stream to avoid synchronization drift. The receiving-end detects the signal synchronization status based on the timestamps and the system clock. However, the timestamp-dependent method has its drawbacks. First, in the sending end, the timestamps are usually added after frame synchronization, format conversion or pre-processing, where the delay derived from these operations are not compensated [18]. Thus, this signal asynchronization in the sending end will take to and always exist in the receiving end. Second, as the sending and receiving ends have different system clocks (the same frequency), the initial delay and frequency offset caused by dynamic environments also lead to signal asynchronization. To solve these shortcomings, researchers have proposed some improvement algorithms. For example, the works in [19,20] utilized the correlation between audio-visual signals for synchronization detection. They extract lip pictures in video frames and then compare them with the features of an audio signal through a deep-learning-based model to determine the synchronization status of audio-visual signals. The limitation of this method is that the video frame must contain the lip region. Yang et al. [21] proposed a watermark-based method to keep the synchronization of the audio-visual signal. However, this method has a disadvantage in that the "watermark" is not well adapted to the video or audio signal when applying conversion, aspect ratio conversion or audio downmixing [18].

From the above analysis, we can make conclusions that:

i. Haptic–visual synchronization plays an important role in HIS. It is worthy of further investigation.
ii. The traditional timestamp-dependent method used in an HIS has some shortcomings. As a result, there is still room for research on the haptic–visual synchronization method.

Thus, In this paper, we propose a first-of-its-kind timestamp-independent synchronization method for haptic–visual signals. Our contributions are summarized as follows.

The sequential correlation between haptic–visual signals. We build a multimedia communication platform with both haptic and visual signals. Based on this platform, we observe a strong correlation between the two signals during haptic-aware interaction. This intrinsic correlation is further utilized to design our synchronization model.

The key sample/frame extraction during haptic–visual interaction. We exploit the statistical features of haptic–visual signals and then develop learning-based methods to extract key samples and key frames in haptic and visual signals, respectively.

The asynchronization detection and removal strategy. Combining the correlation and key samples/frames, we are able to detect and eliminate asynchronization when the registration delay is larger than a threshold. Experimental results with subjective evaluations validate the effectiveness of our method.

2. Motivation

In our opinion, there exists a strong sequential correlation between haptic–visual signals, which can help the judgment of the signal synchronization state without crystal oscillators or timestamps. Inspired by this, we propose a timestamp-independent haptic–visual synchronization model to detect and eliminate asynchronization phenomena in an HIS. In this section, we establish a haptic–visual simulation platform and subsequently confirm the correlation between haptic–visual signals via the platform.

As shown in Figure 1, we use a virtual interaction module to design a haptic–visual interaction scenario where a human user manipulates a virtual ball to push a virtual box. A Geomagic Touch is deployed to connect the real and virtual world: on one hand, it sends the human instructions to the virtual ball; on the other hand, it collects the force feedback of the virtual ball and sends the corresponding signals back to the human user. This haptic interaction is achieved with the kinesthetic signal, which is a major component of haptic information.

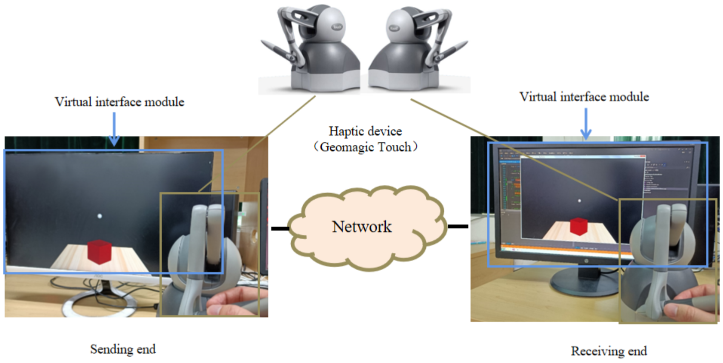


Figure 1. Our simulation platform for haptic–visual signal delivery.

In addition to the haptic signals captured by Geomagic Touch, the sending-end also records the visual contents of the virtual space, resulting in a high-definition video at a resolution of 1920 × 1080. Then the video is compressed by High Efficiency Video Coding (HEVC) and subsequently delivered with haptic signal by the network via User Datagram Protocol (UDP). Finally, the receiving-end combines both haptic and visual signals for a more immersive tele-presence, where another user can watch the scene in real time and also feel the haptic sensing via a haptic device.

The haptic and visual signals should be fully synchronized under normal conditions. Based on this simulation platform, we can observe the sequential correlation between haptic and visual signals. As shown in Figure 2, strong haptic signal fluctuations exist when the virtual hand (i.e., the ball) is on a collision course with another object. When the virtual hand visually touches the box, the force amplitude of the haptic changes simultaneously. As the two objects move closer, the force amplitude is also higher and vice versa. The force amplitude recovers to a constant when all objects are detached. These changes are also intuitive to the human users when operating a haptic-aware handle.

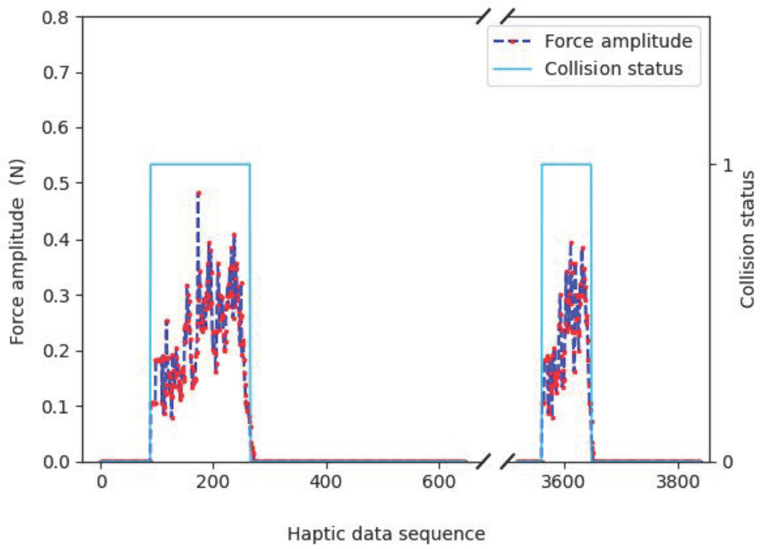
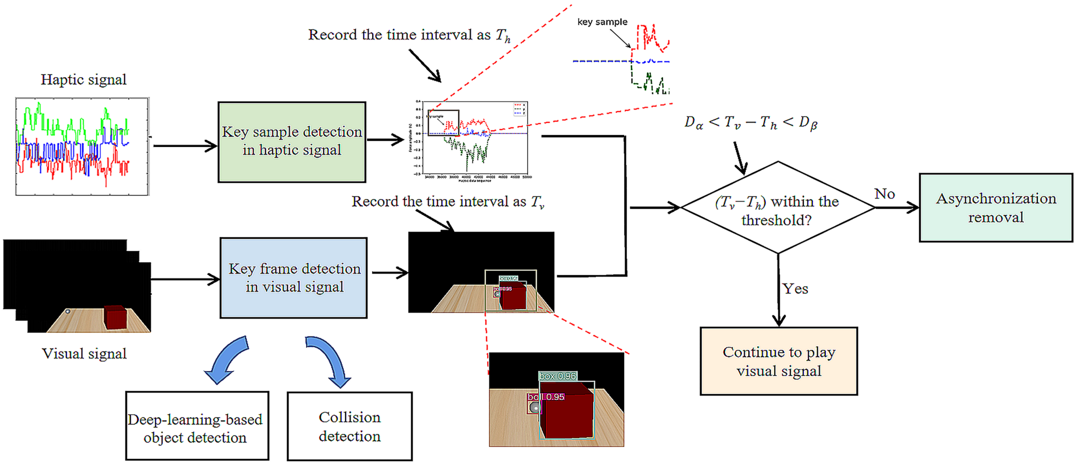


Figure 2. An example of haptic–visual correlations.

This intrinsic correlation inspires us to design a synchronization strategy. A sharp increase of force amplitude indicates a collision between the virtual hand and another object, while a sharp decrease implies a detachment between objects. If these deductions are inconsistent with the machine vision, we can conclude that there exists an asynchronization between haptic and visual signals and thus change the signal flows.

3. Proposed Method

Based on the above analysis, we propose the timestamp-independent synchronization method as shown in Figure 3. First, we extract the key samples in the haptic signal where the amplitude is intensively increased from near zero. Second, we extract the key frames in the visual signal where the visual collision happens. Third, we compare the time intervals of these key samples/frames to detect asynchronization phenomena. If a pair of time intervals (namely T_h and T_v) have a large difference, the haptic–visual asynchronization is found and further fixed. Note that here the object collision frequencies are low in the real world; therefore, we can easily identify different pairs of time intervals. In the following subsections, the key sample detection, key frame detection, threshold selection, asynchronization removal and the overall method are elaborated, respectively.



Record the time interval as T_h
key sample
Haptic signal
Key sample detection in haptic signal
$D_\alpha < T_v - T_h < D_\beta$
Record the time interval as T_v
(T_v-T_h) within the threshold?
No
Asynchronization removal
Yes
Continue to play visual signal
Key frame detection in visual signal
Visual signal
Deep-learning-based object detection
Collision detection
box 0.95

Figure 3. The flowchart of our proposed method.

3.1. Key Sample Detection in the Haptic Signal

For the haptic signal, the key samples are easily obtained for it consists of three one-dimensional signals (in x-axis, y-axis and z-axis). A sharp increase of force amplitude is found when its difference in any dimension is larger than a threshold (namely F_{th}). Through observations on a large number of samples, we found that the fluctuations of force amplitudes during non-collision are always below 0.01, and the force amplitudes of key samples are always above 0.07. Therefore, the F_{th} is empirically set as 0.05 in our work.

An example of this step is shown in Figure 4. An operation with force signals in three dimensions is presented, where all sharp increases are successfully detected and labeled as key samples. Correspondingly, their time intervals (i.e., T_h) are recorded for further comparison.

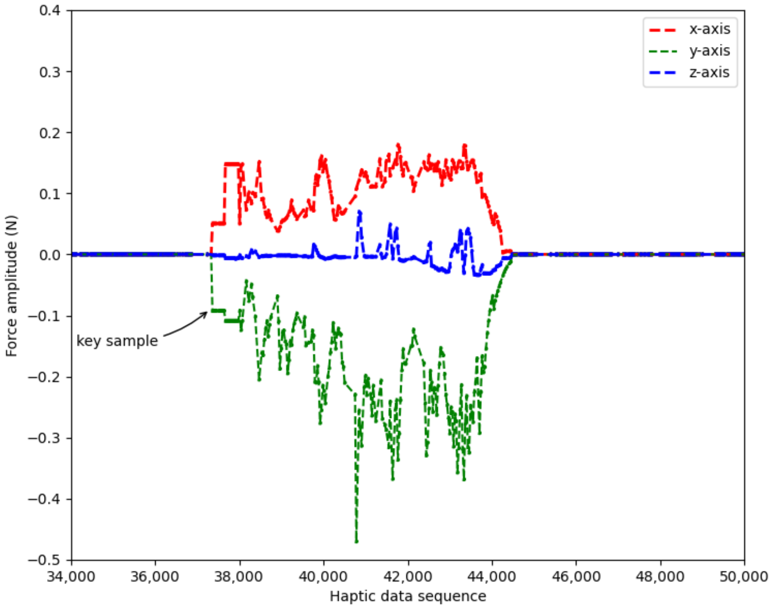


Figure 4. An example of key sample detection.

3.2. Key Frame Detection in the Visual Signal

The objective of key frame detection is to find the time intervals when the virtual hand touches the box. Essentially, it consists of two modules: object detection and collision detection. The first module identifies all objects, while the second module determines whether object collision occurs. Both modules are achieved by computer vision methods.

3.2.1. Object Detection

The commonly-used object detection algorithms are R-CNN [22], SPPNet [23], Fast R-CNN [24], Faster R-CNN [25], SSD [26] and YOLO [27,28]. Considering the efficiency, R-CNN, SPPNet and Fast R-CNN are not suitable for our scenario. Moreover, in our work, small object recognition, in which the performances of the Faster RCNN and SSD are not good enough, is needed. With a deep network, the YOLO network extracts the deep features of different objects and scenarios, thereby achieving object recognition with high accuracy. Consequently, we employ the V3 of YOLO network in our method [28].

We established our image database for training the YOLO V3 network. We acquired 1000 images from visual signals with an image size of 1600 × 900 pixels. Then the images were labeled via a label-making tool (the application software of labelImg). We used a rectangle to bound the balls in the images and labeled them as "ball" and accordingly, bound the boxes and labeled them as "box". All the labels were saved with xml files for using during training. The 800 images in this database are employed as the training set and the other 200 images are the test set.

The loss function plays an important role in the YOLO network. In this work, the position information of the ball and box is the target of the network. Therefore, the target's error of center coordinate in the form of squared difference is first taken into account in the loss function; then, to obtain the accurate bounding rectangle, the wide and high coordinate error in the form of cross-entropy is utilized; finally, as the detection of multiple categories of targets (ball and box) are involved, the category error in the form of cross-entropy must be considered. Hence, the loss function used in this work is:

$$\begin{aligned} Loss = \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+ \\ \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2]- \\ \sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}[\hat{C}_i\log(C_i)+(1-\hat{C}_i)\log(1-C_i)]- \\ \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}[\hat{C}_i\log(C_i)+(1-\hat{C}_i)\log(1-C_i)]- \\ \sum_{i=0}^{S^2} I_{ij}^{obj}\sum_{c\in classes}[\hat{P}_i\log(P_i)+(1-\hat{P}_i)\log(1-P_i)] \end{aligned} \tag{1}$$

where the first row indicates the error of the center coordinates, S represents the grid size, B represents the bounding rectangle. I_{ij}^{obj} denotes whether targets are in the rectangle, and its value is one if there is a target in the bounding rectangle at grid (i, j), and zero vice versa. Here, x_i and y_i represent the true center coordinates; $\hat{x}_i$ and $\hat{y}_i$ represent the predicted center coordinates.

The second row represents the error of the width and height of the predicted rectangle in which w_i and h_i represent the true width and height and $\hat{w}_i$ and $\hat{h}_i$ represent the predicted width and height. The third and fourth rows indicate the error of the confidence level, where C_i denotes the true confidence level, and $\hat{C}_i$ denotes the predicted confidence level.

The fifth row denotes the error of classification, where P_i and $\hat{P}_i$ denote the true and the predicted categories, respectively; λ_{coord} and λ_{noobj} are the weights which will be trained as hyperparameters of the network.

The main hyperparameters used in training are set as shown in Table 1. Among them, the learning rate is set as cosine decay as follows:

$$lr = \left\{\frac{1}{2}\times[1+cos(N_{trained}\times\frac{\pi}{N_{epoch}})]\times 0.95+0.05\right\}\times 10^{-2}, \tag{2}$$

where $N_{trained}$ denotes the number of epochs already trained, and N_{epoch} denotes the total number of training epochs.

Table 1. The hyperparameter settings in model training.

Epoch	Batchsize	λ_{coord}	λ_{noobj}	Learning Rate
300	16	0.5	0.5	cosine decay

With this method, the training module has a larger learning rate at the beginning to accelerate the training speed, and then the learning rate decreases with the increasing number of training epochs to more easily find the optimal solution.

After training, an example of a recognition result is shown in Figure 5 in which the virtual hand (i.e., the ball) and the box are detected, with their borders labeled by rectangular frames.

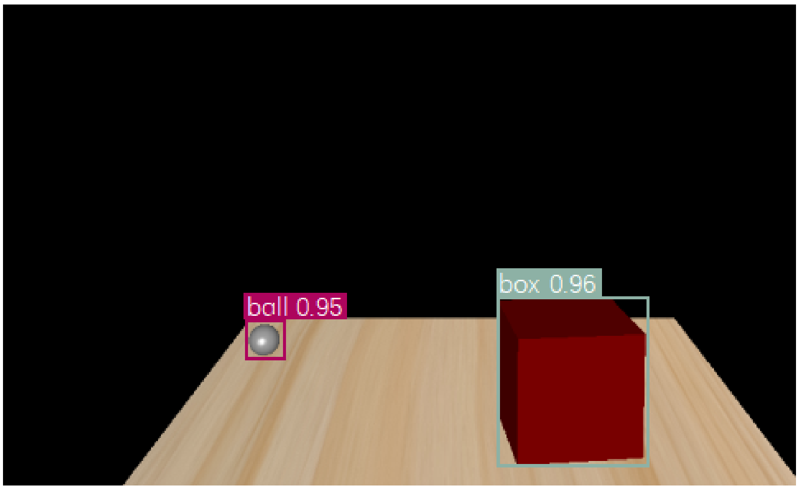


Figure 5. An example of object detection.

3.2.2. Collision Detection

We determine whether a collision happens based on the aforementioned rectangular frames. Let (X_1, Y_1) and (X_2, Y_2) denote the top-left locations of the virtual hand (i.e., the ball) and any object as the target in the 2D space, and (H_1, W_1) and (H_2, W_2) denote the sizes of the corresponding rectangular frames, the condition of no collision is:

$$(Y_1 + H_1 > Y_2)||(X_1 + W_1 < X_2)||(Y_1 < Y_2 + H_2)||(X_1 > X_2 + W_2). \tag{3}$$

Otherwise, the collision of objects is found. At the time of collision found, we extract the corresponding video frame as the key frame of the visual signal and record the time interval as T_v, which is further utilized for asynchronization detection.

3.3. The Synchronization Threshold

During haptic–visual delivery and playback, we can easily identify each key sample/frame pair considering the corresponding time intervals are usually very close to each other. For a pair of time intervals T_h and T_v, their difference is set as a criterion of haptic–visual asynchronization. A synchronization of signals is guaranteed if:

$$D_\alpha < (T_v - T_h) < D_\beta, \tag{4}$$

where D_α and D_β refer to the lower and upper bound of the perception threshold.

As results from a subjective test can be more consistent with users' perception experience, we designed a subjective test to determine D_α and D_β. Our test strictly follows the subjective test manual ITU-R BT.500 [29] with the following steps. First, we recruited 21 subjects without prior knowledge of haptic coding or delivery. Then, we used the two-alternative force choice method to perform the test. Each session of the test consisted of two randomly presented haptic–visual segments: with and without delay. The delay can be negative or positive with a range from -100 ms to 100 ms with an interval of 20ms. Each subject was asked to choose one segment where he/she could not feel delay between the two. Finally, for each session, the probability of correct choices, which is obtained by Equation (5), is recorded.

$$p_i = \frac{n_i}{N}, \tag{5}$$

where n_i denotes the number of subjects who have made a correct choice in the i-th delay, and N denotes the total number of subjects.

As shown in Figure 6, the probability of correct choices is around 0.5 when the delay of visual signals ranges from −60 ms to 80 ms. In other words, the human users cannot perceive the difference between delayed and non-delayed signals in this range. Therefore, we set the threshold of synchronization as $D_\alpha = -60$ ms, $D_\beta = 80$ ms.

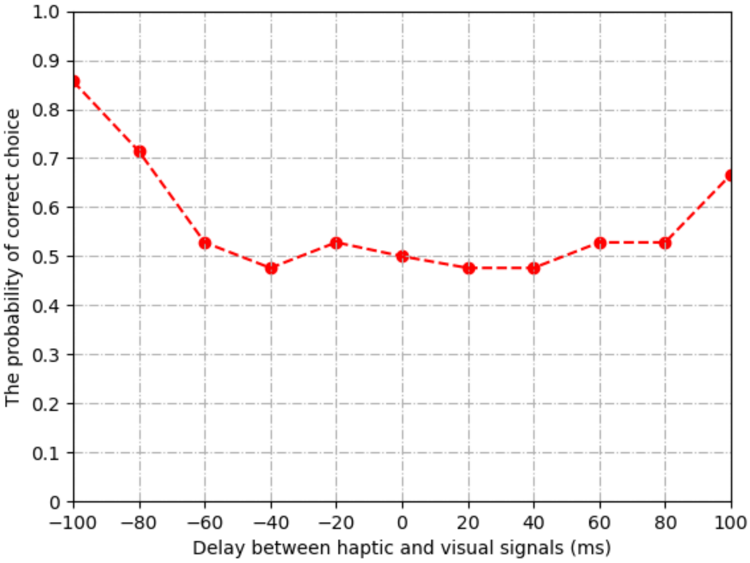


Figure 6. Subjective result of synchronization threshold.

3.4. *Asynchronization Removal*

To adjust the signal stream and remove asynchronization phenomena, a general method is to select a main stream and set the remaining as auxiliary streams. When asynchronization occurs, all auxiliary streams are adjusted to be synchronized with the main stream. As reported in [30], the human perception of haptic signals is very sensitive in that only haptic signals above 1 kHz provide smooth experience to users. This frequency is significantly higher than visual signals. Based on this fact, we utilize the haptic signal and the visual signal as the main stream and the auxiliary stream, respectively. For synchronization, the visual signal is moved to be consistent with the haptic signal.

In a multimedia communication system, the receiving-end usually sets a buffer zone to cache all multimedia data for a smooth display of them. Therefore, if the visual signal is delayed more than D_α, we will retrieve the correct video frame from the buffer zone. otherwise, if the visual signal is ahead by D_β, we will repeat the current frame until haptic–visual synchronization. Through this method, we are able to remove all asynchronization phenomena during haptic–visual delivery and playback.

3.5. *The Overall Method*

By summarizing Sections 3.1–3.4, the detailed steps of our method are presented as follows.

Step 1. Initialization. Set a buffer zone at the receiving end to cache haptic–visual data. Start the haptic–visual data delivery and playback. Go to Step 2.

Step 2. Key sample detection. Keep to detect the key samples of the haptic signals with the method in Section 3.1. If a key sample is found, set the time interval as T_h and go to Step 3.

Step 3. Key frame detection. Use the method in Section 3.2 to detect the corresponding key frames in the buffer and subsequent video of 1 s. If a key frame is found, set the time interval as T_v and go to Step 4; otherwise, the synchronization detection fails, go to Step 2.

Step 4. Asynchronization examination. If Equation (4) of Section 3.3 is true, go to Step 2 to check the following signals; otherwise go to Step 5.

Step 5. Asynchronization removal. Adjust the haptic–visual streams with the method shown in Section 3.4. Go to Step 2 to check the following signals.

4. Experimental Results

To examine the effectiveness of the proposed method, we implement it on the simulation platform shown in Section 2 and conduct both objective and subjective experiments. The frequencies of haptic and visual signals are set as 1000 Hz and 30 Hz, respectively. Due to the lack of a haptic–visual synchronization method, we compare our model with the original case only.

4.1. Estimation Accuracy of Synchronization Delay

The proposed method utilizes the synchronization delay $T_v - T_h$ to determine whether asynchronization happens. Therefore, the estimation accuracy of synchronization delay is critical in our method. We design the following experiment to examine the accuracy.

Based on the simulation platform, we randomly captured 100 haptic–visual clips, with the length of each clip as 30 s. In other words, there exist 30,000 haptic samples and 900 video frames in each clip; in total, 3 million haptic samples and 90,000 video frames exist). For each haptic–visual clip, we add a random delay on the visual signals. The delay is in the range of (−330 ms, 330 ms) where the positive/negative values indicate the visual signal is ahead/behind the haptic signal. At the receiving-end, we employ our model to calculate the synchronization delay (namely $\hat{d}$) and compare it with the "actual" delay (namely d).

The Mean Absolute Error (MAE) and Maximum Absolute Error (MaxAE) are utilized to be assessment metrics. They are calculated by:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left|\hat{d}_i - d_i\right|, \tag{6}$$

$$MaxAE = \max_{i\in\{1,2,3,..M\}}\left|\hat{d}_i - d_i\right|, \tag{7}$$

where M is the total number of samples.

The results are shown in Table 2. From the table, the MAE and MaxAE values are 7.3 ms and 15 ms, respectively. It is noted that the haptic–visual synchronization is unperceivable in (−60 ms, 80 ms), where the ratio of MAE and MaxAE are only 5.2% and 10.7%, respectively. On the other hand, the frame length of each video frame is $\frac{1}{30}$ Hz = 33.3 ms, which is also significantly larger than the MAE/MaxAE values. Therefore, the estimation accuracy could fulfill the requirement in the practical applications of the haptic–visual system.

Table 2. The estimation accuracy of $T_v - T_h$.

Metrics	MAE (ms)	MaxAE (ms)
Results	7.3	15

4.2. Effectiveness of the Haptic–Visual Synchronization

To evaluate the effectiveness of our synchronization detection and removal method, we examine it on the same dataset presented in Section 4.

At the sending end, after sending random video frames (in the range of (0, 100)), we add a random delay (in the range of (−330 ms, 330 ms) and denoted as t_n) on it. We repeat the above process until all the frames in each clip (totally 100 clips) are sent. Considering that the proposed asynchronization removal method adjusts the visual signal frame-by-frame, the interval of the above random delay is set the same as the frame interval of the visual signal (i.e., 33 ms). Therefore, the delay range of (−330 ms, 330 ms) is equivalent to a delay random number (denoted as d_n) of video frames in the range of (−10, 10). Taking a clip (900 frames) as example, the random numbers generated in the experiment are shown in Table 3. In the table, the values in the first column indicate that the visual signal is ahead of the haptic signal 7 × 33 = 231 ms, and the delay status lasts for 19 × 33 = 627 ms. The above random delay in the experiment is also intuitively shown in Figure 7 in which the vertical axis indicates the delay between visual and haptic signals and the horizontal axis indicates the order of the visual signal. From the figure, the delays are random and representative to evaluate our method.

Table 3. An example of random delay in the experiment.

d_n	**7**	**−7**	**8**	**−8**	**8**	**9**	**−1**	**0**	**5**	**−8**	**−8**	**2**	**0**	**1**	**−1**	**−7**
t_n	19	18	95	17	56	65	82	46	69	96	47	86	36	99	14	55

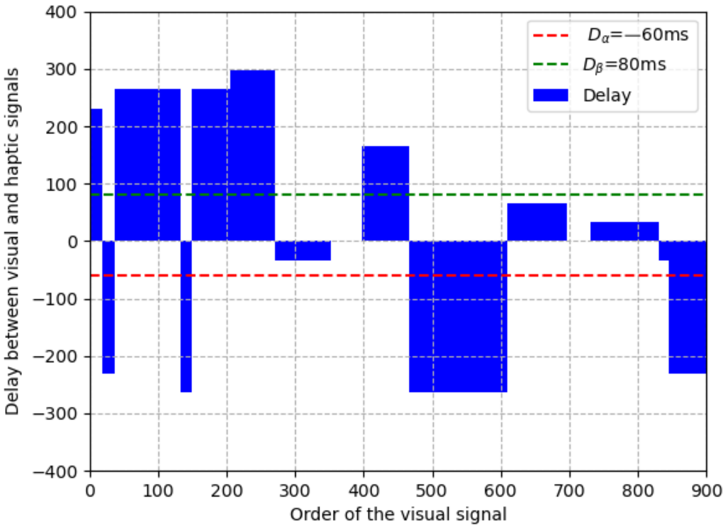


Figure 7. An example of random delay in the experiment.

At the receiving end, we compare the probabilities of successful synchronization with and without our method. The results are presented in Table 4. By using our model, the average probability of synchronization increases from 25.3% to 89.2%. It should be pointed out that our synchronization method is executed frame-by-frame. If the haptic–visual delay is larger than one frame, the signal is kept asynchronized during the synchronization process. That is the reason why there are still 10.8% signals asynchronized in Table 4. Even at this scenario with severe fluctuations, our method still achieves a high probability of 89.2%, which reveals the effectiveness and robustness of our method in haptic–visual synchronization. The utilization of our model guarantees the signal synchronization in most cases, thereby greatly improving the system performance of haptic–visual interaction.

Table 4. Probabilities of synchronization with and without our method.

	Without Our Method	**With Our Method**
Probabilities	25.3%	89.2%

4.3. Subjective Test on User Experience

In addition to objective evaluation, we also conducted a subjective test to evaluate the improvement of the user experience with our model. As mentioned in Section 1, the signal asynchronization is a critical factor to influence the user experience in haptic–visual interaction. Therefore, the improvement of user experience can be taken as circumstantial evidence of the effectiveness of our model.

We recruited 23 subjects to participate in this test, where all haptic–visual sequences are also the same to those in Section 4.1. The subjects' ages ranged from 17 to 26, and they have no exposure to the haptic-based system. To calculate the correlations, we introduce the delays that are evenly distributed from −10 to 10 frames (that is, ranged from −333 ms to

333 ms with the interval of 33.3 ms) and occasionally utilize the proposed synchronization method at the receiving end. However, whether or not we are using the synchronization is unknown for all subjects. As a result, a subject scores his/her experience based on real feelings and experiences. All scores are between 0 and 10 and their averaged value, the Mean Opinion Score (MOS), represents the average perceptions of human users.

The collected subjective test results were pre-processed to remove outliers based on the ITU subjective test regulations. We calculated the correlations, including the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SROCC) [31], between each subject's score and the MOS. The results are shown in Figure 8. According to ITU-R BT.500 [29], a subject's score is considered as an outlier if the correlation between his/her score and the MOS is less than 0.7. Therefore, from Figure 8, the 12th and 18th subjects are considered as outliers and subsequently excluded in the final results.

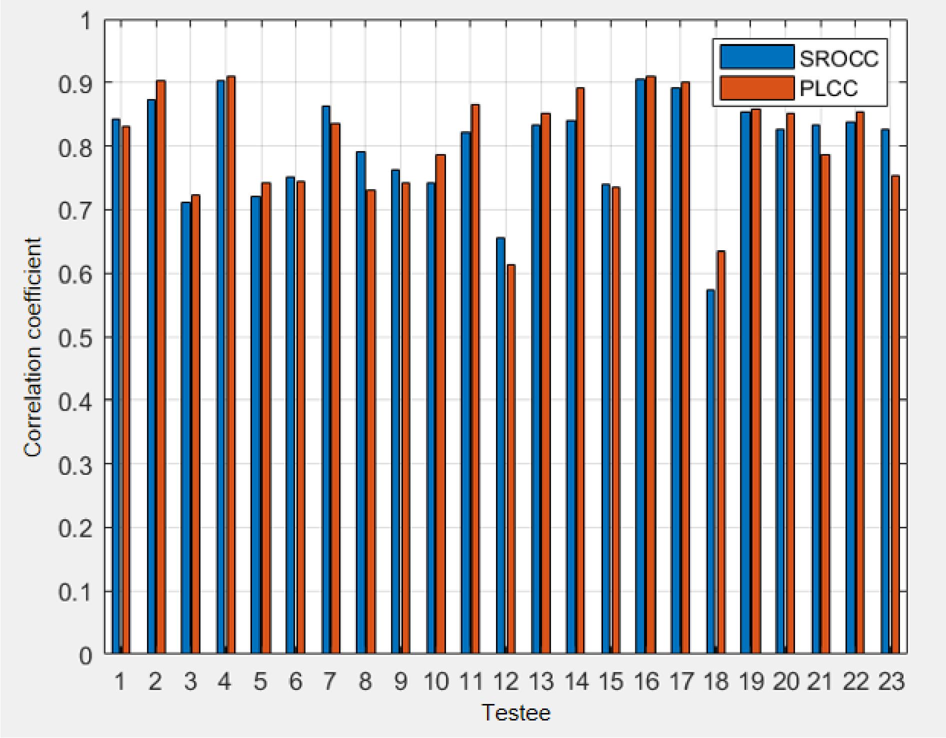


Figure 8. The correlations between each subject and the MOS.

The scores of the remaining 21 subjects were further examined by data saturation validation [32]. Due to the randomness of user scores, insufficient subjects would lead to inaccurate MOS values. To check whether the subjects are enough, data saturation validation was proposed. For a subjective test with K subjects, it randomly selects $k = 1, 2, \ldots, K$ subjects to calculate the correlation between their averaged score and the MOS. If the correlation value converges to one as k increases, the subjects are considered sufficient. In our test, this correlation value is very close to 1 with $k = 13$ subjects, as shown in Figure 9. Therefore, the remaining 21 subjects are sufficient to represent the averaged opinion of human users.

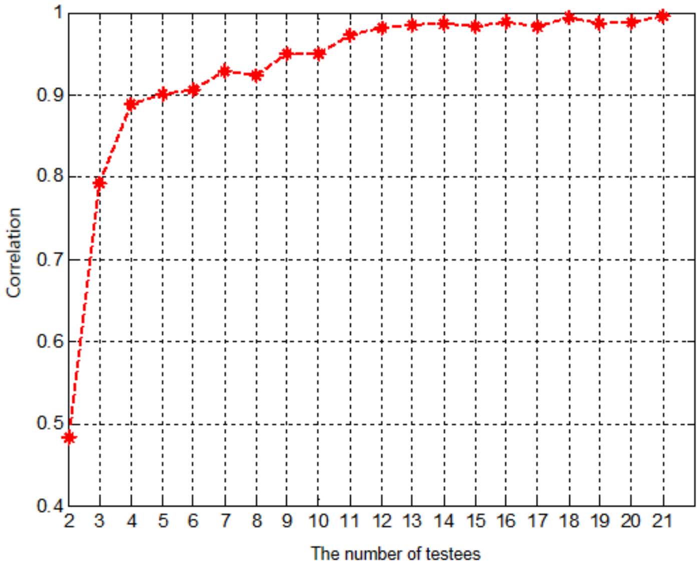


Figure 9. The data saturation validation in our test.

Figure 10 shows the MOS values under different delay settings. Two settings are compared: receiving end with and without our method. In the central part of curves (i.e., −33~66 ms), the delays are unperceivable to human users; thus the two settings achieve very similar MOS values. As the absolute value of delay gets larger, the difference between the two settings becomes more significant. In extreme cases (i.e., ±330 ms), our synchronization method improves the MOS values by around four, which shows the high capability of anti-interference under severe network conditions. On average, the MOS value is increased by 1.6169, with MOS variation decreased by 3.1315. This fact demonstrates the significant improvement of our synchronization method that is agreed by the majority of human users. In conclusion, the proposed method can guarantee the user experience in case of haptic–visual asynchronization.

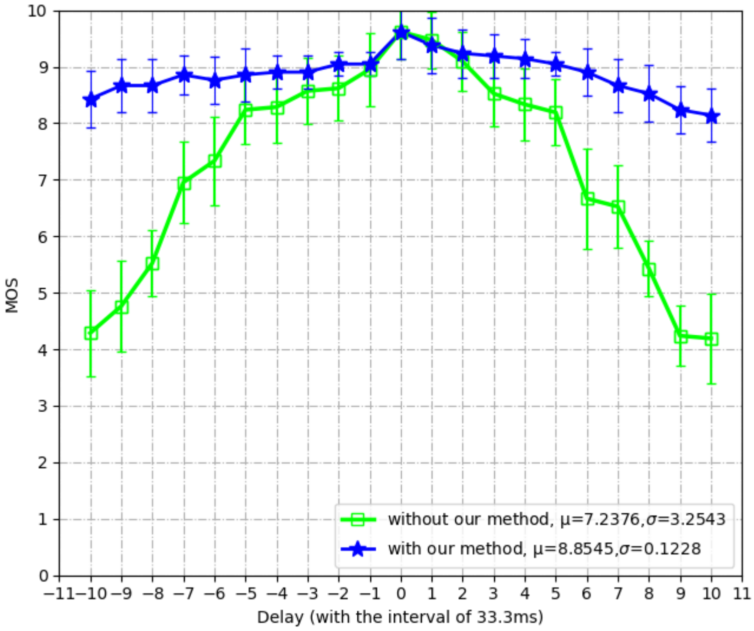


Figure 10. The subjective improvements with our method.

5. Conclusions

In this paper, we explore the haptic–visual correlations in a haptic-aware interaction system. Based on the observations, we propose a timestamp-independent synchronization method for haptic–visual signals, which consists of haptic signal analysis, learning-based vision analysis, perception-based thresholding and an overall method for asynchronization detection and removal. It should be pointed out that the example of virtual hand (i.e., the ball) and target (i.e., the box) can be extended to more types of objects with retrained models. Therefore, our model is still applicable in more general scenarios. To our best knowledge, this is the very first attempt to design a haptic-aware multimedia synchronization model by considering the special characteristics of haptic interaction. It can also be utilized as a reference to design new synchronization models for emerging sensorial media such as olfactory signals. We envision a more widespread use of multiple sensorial media that benefits the immersive user experience in the foreseeable future.

Author Contributions: Conceptualization, Y.X. and T.Z.; methodology, Y.X. and L.H.; software, Y.X. and L.H.; validation, Y.X. and T.Z.; formal analysis, L.L.; investigation, Y.F. and L.L.; resources, Y.X. and T.Z.; data curation, Y.X. and L.H.; writing—original draft preparation, Y.X. and L.H.; writing—review and editing, Y.X. and L.H.; visualization, Y.X. and L.H.; supervision, Y.X. and T.Z.; project administration, Y.X. and T.Z.; funding acquisition, Y.X. and T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work were supported by the National Natural Science Foundation of China (No. 62171134) and Foundation for Middle-aged and Young Educational Committee of Fujian Province (No. JAT200024).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aijaz, A.; Dohler, M.; Aghvami, A.H.; Friderikos, V.;Frodigh, M. Realizing the Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks. *IEEE Wirel. Commun.* **2017**, *24*, 82–89. [CrossRef]
2. Antonakoglou, K.; Xu, X.; Steinbach, E.; Mahmoodi, T. Toward Haptic Communications Over the 5G Tactile Internet. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 3034–3059. [CrossRef]
3. Qiao, Y.; Zheng, Q.; Lin, Y.; Fang, Y.; Xu, Y.; Zhao, T. Haptic Communication: Toward 5G Tactile Internet. In Proceedings of the 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), Fuzhou, China, 13–16 December 2020; pp. 1–3.
4. Steinbach, E.; Strese, M.; Eid, M.; Liu, X.; Bhardwaj, A.; Liu, Q.; Al-Ja'afreh, M.; Mahmoodi, T.; Hassen, R.; El Saddik, A.; et al. Haptic Codecs for the Tactile Internet. *Proc. IEEE* **2019**, *107*, 447–470. [CrossRef]
5. Xu, Y.; Huang, Y. Chen, W.; Xue, H.; Zhao, T. Error Resilience of Haptic Data in Interactive Systems. In Proceedings of the 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 23–25 October 2019; pp. 1–6.
6. Bortone, I.; Leonardis, D.; Mastronicola, N.; Crecchi, A.; Bonfiglio, L.; Procopio, C.; Solazzi, M.; Frisoli, A. Wearable Haptics and Immersive Virtual Reality Rehabilitation Training in Children With Neuromotor Impairments. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 1469–1478. [CrossRef] [PubMed]
7. Zhou H.; Wei, L.; Cao, R.; Hanoun, S.; Bhatti, A.; Tai, Y.; Nahavandi, S. The Study of Using Eye Movements to Control the Laparoscope Under a Haptically-Enabled Laparoscopic Surgery Simulation Environment. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018; pp. 3022–3026.
8. Chen, Y.; Zhu, J.; Xu, M.; Zhang, H.; Tang, X.; Dong, E. Application of Haptic Virtual Fixtures on Hot-Line Work Robot-Assisted Manipulation. In *Intelligent Robotics and Applications*; Yu, H., Liu, J., Liu, L., Ju, Z., Liu, Y., Zhou, D., Eds.; Publishing House: Hefei, China, 2019; pp. 221–232.
9. Durai, V.S.I.; Arjunan, R.; Manivannan, M. The Effect of Audio and Visual Modality Based CPR Skill Training with Haptics Feedback in VR. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; pp. 910–911.
10. Decré, G.B.; Cloonan, C. A touch of gloss: Haptic perception of packaging and consumers' reactions. *J. Prod. Brand Manag.* **2019**, *11743*, 117–132. [CrossRef]

11. Wong, H.; Kuan, W.; Chan, A.; Omamalin, S.; Yap, K.; Ding, A.; Soh, M.; Rahim, A. Deformation and Friction: 3D Haptic Asset Enhancement in e-Commerce for the Visually Impaired. *Haptic Interact. AsiaHaptics* **2018**, *535*, 256–261.
12. Lisini Baldi, T.; Scheggi, S.; Aggravi, M.; Prattichizzo, D. Haptic Guidance in Dynamic Environments Using Optimal Reciprocal Collision Avoidance. *IEEE Robot. Autom. Lett.* **2018**, *3*, 265–272. [CrossRef]
13. Da Fonseca, V.P.; Monteiro Rocha Lima, B.; Alves de Oliveira, T.E.; Zhu, Q.; Groza, V.Z.; Petriu, E.M. In-Hand Telemanipulation Using a Robotic Hand and Biology-Inspired Haptic Sensing. In Proceedings of the 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Istanbul, Turkey, 26–28 June 2019; pp. 1–6.
14. Li, S.; Rameshwar, R.; Votta, A.M.; Onal, C.D. Intuitive Control of a Robotic Arm and Hand System With Pneumatic Haptic Feedback. *IEEE Robot. Autom. Lett.* **2019**, *4*, 4424–4430. [CrossRef]
15. Zeng, Q.; Ishibashi, Y.; Fukushima, N.; Sugawara, S.; Psannis, K.E. Influences of inter-stream synchronization errors among haptic media, sound, and video on quality of experience in networked ensemble. In Proceedings of the 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE), Tokyo, Japan, 1–4 October 2013; pp. 466–470.
16. Aung, S.T.; Ishibashi, Y.; Mya, K.T.; Watanabe, H.; Huang, P. Influences of Network Delay on Cooperative Work in Networked Virtual Environment with Haptics. In Proceedings of the 2020 IEEE Region 10 Conference (TENCON), Osaka, Japan, 16–19 November 2020; pp. 1266–1271.
17. El-Helaly, M.; Amer, A. Synchronization of Processed Audio-Video Signals using Time-Stamps. In Proceedings of the 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 16 September–19 October 2007; pp. VI-193–VI-196.
18. Staelens, N.; Meulenaere, J.D.; Bleumers, L.; Wallendael, G.V.; Cock, J.D.; Geeraert, K.; Vercammen, N.; Broeck, W.; Vermeulen, B.; Walle, R. Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. *Multimed. Syst.* **2012**, *18*, 445–457. [CrossRef]
19. Kikuchi, T.; Ozasa, Y. Watch, Listen Once, and Sync: Audio-Visual Synchronization With Multi-Modal Regression Cnn. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3036–3040.
20. Torfi, A.; Iranmanesh, S.M.; Nasrabadi, N.; Dawson, J. 3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition. *IEEE Access* **2017**, *5*, 22081–22091. [CrossRef]
21. Yang, M.; Bourbakis, N.; Chen, Z.; Trifas, M. An Efficient Audio-Video Synchronization Methodology. In Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 767–770.
22. Yang, L.; Song, Q.; Wang, Z.; Hu, M.; Liu, C. Hier R-CNN: Instance-Level Human Parts Detection and A New Benchmark. *IEEE Trans. Image Process.* **2021**, *30*, 39–54. [CrossRef] [PubMed]
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
24. Ullah, A.; Xie, H.; Farooq, M.O.; Sun, Z. Pedestrian Detection in Infrared Images Using Fast RCNN. In Proceedings of the 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), Xi'an, China, 7–10 November 2018; pp. 1–6.
25. Gomzales, R.; Machacuay, J.; Rotta, P.; Chinguel, C. Faster R-CNN with a cross-validation approach to object detection in radar images. In Proceedings of the 2021 IEEE International Conference on Aerospace and Signal Processing(INCAS), Lima, Peru, 28–30 November 2021; pp. 1–4.
26. Ahmad, T.; Chen, X.; Saqlain, A.; Ma, Y. EDF-SSD: An Improved Feature Fused SSD for Object Detection. In Proceedings of the 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 24–26 April 2021; pp. 469–473.
27. Wang, Z.; Xie, K.; Zhang, X.; Chen, H.; Wen, C.; He, J. Small-Object Detection Based on YOLO and Dense Block via Image Super-Resolution. *IEEE Access* **2021**, *9*, 56416–56429. [CrossRef]
28. Chen, H.; He, Z.; Shi, B.; Zhong, T. Research on Recognition Method of Electrical Components Based on YOLO V3. *IEEE Access* **2019**, *7*, 157818–157829. [CrossRef]
29. International Telecommunication Union. *Methodology for the Subjective Assessment of the Quality of Television Pictures*; International Telecommunication Union: Geneva, Switzerland, 2002.
30. Huang, P.; Sithu, M.; Ishibashi, Y. Media Synchronization in Networked Multisensory Applications with Haptics. In *MediaSync*; Montagud, M., Cesar, P., Boronat, F., Jansen, J., Eds.; Publishing House: Grao de Gandia, Spain; Amsterdam, The Netherlands, 2018; pp. 295–317.
31. Thirumalai, C.; Chandhini, S.A.; Vaishnavi, M. Analysing the concrete compressive strength using Pearson and Spearman. In Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; pp. 215–218.
32. Yang, H.; Bao, B.; Guo, H.; Jiang, Y.; Zhang, J. Spearman Correlation Coefficient Abnormal Behavior Monitoring Technology Based on RNN in 5G Network for Smart City. In Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15–19 June 2020; pp. 1440–1442.