Video Content Analysis for Compliance Audit in Finance and Security Industry

BIN LIU¹, MINGYU WU¹, MINZE TAO², QIN WANG², LUYE HE², GUOLIANG SHEN³, KAI CHEN¹, (Member, IEEE), AND JUNCHI YAN¹,4, (Member, IEEE)

¹School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
²E-Capital Transfer Company Ltd., Shanghai 201206, China
³Zhejiang Zheneng Natural Gas Operation Company, Ltd., Hangzhou 310000, China
⁴Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China

Corresponding author: Junchi Yan (yanjunchi@sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100704 and Grant 2016YFB1001003, in part by the NSFC under Grant 61972250 and Grant U19B2035, and in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

ABSTRACT The quality and accessibility of modern financial service have been quickly and dramatically improved, which benefits from the fast development of information technology. It has also witnessed the trend for applying artificial intelligence related technology, especially machine learning to the finance and security industry ranging from face recognition to fraud detection. In particular, deep neural networks have proven to be far superior to traditional algorithms in various application scenarios of computer vision. In this paper, we propose a deep learning-based video analysis system for automated compliance audit in stock brokerage, which in general consists of five modules here: 1) Video tampering and integrity detection; 2) Objects of interest localization and association; 3) Analysis of presence and departure of personnel in a video; 4) Face image quality assessment; and 5) Signature action positioning. To the best of our knowledge, this is the first work that introduces remote automated compliance audit system for the dual-recorded video in finance and security industry. The experimental results suggest our system can identify most of the potential non-compliant videos and has greatly improved the working efficiency of the auditors and reduced human labor costs. The collected dataset in our experiment will be released with this paper.

INDEX TERMS Compliance audit, dual record, video content analysis, deep learning, object localization, action recognition.

I. INTRODUCTION

The Measures for the Suitability Management of Securities and Futures Investors (Hereinafter refer to as the Measures) of China has come into force in July 27, 2017. As the first special regulation of investor protection in China’s securities and futures market, this act is an important foundation law of the capital market. The act will regulate the behavior of the brokers and also provide evidence for the settlement of disputes in the future.

The Measures requires that when a staff of a securities or insurance company sells financial products and signs cooperation documents, the company should be responsible for recording video and audio of the business processing that complies with certain rules for verification of possible disputes which was called dual-record. The dual-record video usually consists of two persons in a specific scenario which includes a staff of securities or insurance company and a client. From the visual quality perspective, the video should meet the following conditions before it can be called a compliant video:

- The video must be complete and continuous and can’t be modified.
- The financial product salesman should wear work uniform and work name tag during video recording and the company nameplate/logo should appear in the recording background.
- The faces of both parties should be unmasked and clearly identified.
- Neither the server (staff) nor the client (customer) can leave the scene during the recording.

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.
The video should contain the entire process of document signing.

In fact, to our best knowledge, currently the inspection and review tasks of the dual-recorded video is manually performed to ensure that the video complies with the corresponding rules. However, there are several problems with manual audit: 1) the length of the video is always longer than 15 minutes, which can be a very tedious work for manual review; 2) video tampering and integrity rely on frame-by-frame analysis, which is almost impossible for manual check; 3) auditing is a multi-tasking process that requires repeated analysis and review. The above problems may cause such situations that most of the videos cannot be carefully reviewed before they are stored and when a dispute occurs, the stored videos may fail to play their due role as they do not meet the dual-recorded requirements. If accurate automated compliance analysis of the video content can be performed in a real-time while the video is being archived, it should be well avoided.

As the deep convolution neural networks (DCNN) is powerful in learning robust and high-level semantic feature representations of the image/video sequence, it has fueled great strides in a variety of computer vision tasks. In image classification, despite some early successes of convolution neural network (CNN) (e.g. [1], [2]), the DCNN brings us to the new era of computer vision filed (e.g. [3]–[5]). The most significant advance was achieved in the Large Scale Visual Recognition Challenge 2012 [6], Krizhevsky et al. [7] used the DCNN which was called AlexNet to train on about 1.2 million images, and finally got the record-breaking result. Since then, DCNN based methods have dominated most of the tracks in the ILSVRC; Object detection is one of the most fundamental problem in computer vision, it seeks to locate object instances from a large number of predefined categories in natural images. Girshick et al. took the lead to apply DCNN in object detection in 2014 by proposing the Regions with CNN features (RCNN) for object detection [8]. Since then, object detection started to evolve at an unprecedented speed. Girshick also proposed Fast RCNN [9] that addresses some of the disadvantages of RCNN, while improving on their detection speed and quality. Although Fast RCNN significantly speeds up the detection process, it still relies on external region proposals, whose computation is exposed as the new speed bottleneck in Fast RCNN. To solve this problem, the Faster RCNN framework proposed by Ren et al. [10] offered an efficient and accurate Region Proposal Network (RPN) for generating region proposals. RCNN-family methods are typically slower because of large amount of region proposals which should be classified and regressed on by one later, and in order to improve the speed, single-stage detectors, such as YOLOv3 [11] and SSD [12], that treat object detection as a simple regression problem were proposed; Recently, the Siamese network based trackers have drawn much attention in the community. These Siamese trackers formulate the visual object tracking problem as learning a general similarity map by cross-correlation between the feature representations learned for the target template and the search region. SiamPRN [13] introduces a region proposal network after the Siamese network, thus formulating the tracking as a one-shot local detection task; In face detection, MTCNN [14] proposes a deep cascaded multi-task framework which exploits the inherent correlation between detection and alignment to boost up their performance, and various other tasks.

Video content analysis (VCA) deals with the extraction of structural metadata from raw video to be used as components for further processing in applications such as summarization, classification or event detection, and the goal is to extract structural and semantic content automatically for handling the problem of detecting abnormal events [15]–[17], which is a key subject in video surveillance, where the DCNN has also been widely used to determine anomaly of the scene. From the perspective of video content analysis, we focus on detecting the following non-compliance behaviors: a. the video is incomplete and tampering such as frame insertion or frame deletion; b. the uniformity of the staff is not according to the Measures, and the recording site does not contain identifiers such as company logo or brand which can be used to identify the company; c. it is impracticable to identify the participants by their low quality faces in the video sequence; d. the process of signing a document cannot be detected and pinpointed in the video. Based on deep learning technology, this paper effectively integrates the higher-precision and computationally efficient detection, recognition, and tracking deep learning neural networks, as well as basic methods in the traditional digital image processing field.

The contributions of our paper are summarized as follows:

- This is the first work that introduces remote automated compliance audit system for dual-recorded video in the financial and securities industry.
- We proposed a series of practical solutions for compliance detection, which include the joint discrimination mechanism of Mean Structural Similarity Index (MSSIM) [18], luminance compensation and deep content analysis in special scenarios is proposed for video integrity detection in complex scenarios; An elaborately designed object association scheme which can establish the relationship bonding the objects and corresponding person, moreover scoring mechanism which can represent facial pose, size, sharpness, and feature attributes in face quality assessment were used in this particular scenario; 3D CNN [19] is employed to recognize signature action, and we used Siamese-RPN to filter background noise in video sequence which greatly improves the accuracy.
- We built a dataset for object detection in the financial scenario by image search engines such as Google image, Baidu image search engine, etc., and we will make part of it public available together with the publication of this paper.
The remainder of this paper is organized as follows: Section II gives a brief introduction to related work. In Section III, we describe the related methodologies for each module of this automated system. In Section IV the implementation details and case study are presented. Finally, Section V concludes the paper with a summary of findings and future works.

II. RELATED WORKS

Video content analysis (VCA) is the capability of automatically analyzing video for specific tasks based on video sequence instead of single image. This technical capability is widely used in the security industry for better sensing of the behind situation. The video data generated by IP cameras are used for data analysis to process, categorize and analyze the objects and activities captured. In our proposed system, human action and motion recognition is important part which has a wide range of applications, such as intelligent video surveillance and environmental home monitoring [20], video storage and retrieval [21], intelligent human–machine interfaces [22], and identity recognition [23]. Human action and motion recognition cover many research topics, including human detection in video, human pose estimation, human tracking, and analysis and understanding of time series data. The key to good human action recognition is robust human action modeling and feature representation. Unlike feature representation in the image space, the feature representation of human action in video not only describes the appearance of the human in the image space, but must also extract the motion in the image sequence. The problem of feature representation is extended from two-dimensional image space to three-dimensional spatio-temporal space.

Over recent years, there are many works focusing on applying CNN models to the video domain with an aim to learn spatio-temporal patterns which have received increasing attention in video content analysis (VCA). Using the 2D CNN is a straightforward way to conduct video recognition. For example, Simonyan and Zisserman [24] designed a two-stream CNN for RGB input (spatial stream) and optical flow input (temporal stream) respectively. Ji et al. [19] introduced the 3D CNN model that operates on stacked video frames. The 3D CNN utilizes 3D kernels for convolution to learn motion information between adjacent frame in volumes. To effectively handle 3D signals, Sun et al. [25] introduced factorized spatio-temporal convolution networks that factorize the original 3D convolution kernel learning as a sequential process of learning 2D spatial kernels in the lower layer. Wang et al. [26] proposed a spatial-temporal non-local module to capture long-range dependencies.

As you know, there are lots of repetitive tasks in the manual audit. Therefore, our goal is helping to reduce jobs in which AI and machine learning take over boring tasks, while humans spend more time with higher-level tasks. In this paper, we present an automated compliance audit problem in the real business scenario and propose an intelligent audit system based on a set of deep learning technologies to simplify the boring tasks and help the staff to improve their work efficiency. The system integrates object detection, object tracking, face detection and traditional image processing methods together to effectively solve the following problems, such as video tampering and integrity validation, object localization and object association, person-departure detection, face image quality assessment and signature action positioning. We divided the system to five corresponding business modules, see Figure 1, which includes video tampering and integrity analysis; badge, company logo, nameplate and uniform detection; checking if both the staff and the customer left the camera view during recording; face image quality assessment in video sequence; and signature action detection.

FIGURE 1. Architecture of proposed automated dual-recorded video compliance audit system. It contains two parts: algorithm module which is based on deep networks, and business function module.

III. PROPOSED APPROACH FOR VISUAL ANALYSIS BASED COMPLIANCE AUDIT

In this section, we describe the detailed method of each sub-module in our proposed approach, which includes video tampering and integrity detection, object detection and object association, person-departure detection, face image quality assessment and signature action positioning.

A. VIDEO TAMPERING AND INTEGRITY DETECTION

1) VISUAL FEATURE

To detect if the video is integrated and not tampered by others, we use the Structural Similarity Index (SSIM) [27] which is a perceptual metric that quantifies the image quality degradation to detect the modification. SSIM measures the perceptual difference between two similar images, so it is sensitive to image changing in short time. Here in our system, the Mean Structural Similarity Index (MSSIM) [18] is used for measuring the difference between the two images, which is often more effective to capture the local changes of the image.

SSIM compares the similarity between two images x and y in three dimensions: luminance I(x, y), contrast c(x, y), and
structure \( s(x, y) \). Finally, the similarity between \( x \) and \( y \) is a function of these three terms:

\[
SSIM(x, y) = \left( l(x, y) \right)^p \left( c(x, y) \right)^r \left( s(x, y) \right)^t
\]  

(1)

The brightness similarity of the two images is calculated by the following formula:

\[
l(x, y) = \frac{2 \mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}
\]  

(2)

where \( \mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \) is the mean intensity of the image and \( C_1 \) is a constant value to avoid division by zero. The contrast similarity of the two images is calculated by the following formula:

\[
c(x, y) = \frac{2 \sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}
\]  

(3)

where \( \sigma_x = \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} \) is the standard deviation of image \( x \). And the formula of structural similarity is as follows:

\[
s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3}
\]  

(4)

where \( \sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \) is the covariance of \( x \) and \( y \).

In general, SSIM cannot be directly applied to evaluate the similarity of the whole image. One can divide the image into blocks, and then the SSIM of each block is calculated, which are finally averaged to generate the image-level score:

\[
MSSIM(X, Y) = \frac{1}{M} \sum_{j=1}^{m} MSSIM(x_j, y_j)
\]  

(5)

2) DEEP SEMANTIC FEATURE
The MSSIM based methods work fine in most of the common cases, but in some situations where violent lens shaking or strong light noise appear in the scene, which may cause false alarms. Therefore, our approach recent adopts the recent advent of deep residual neural networks [28] which is often pre-trained on ILSVRC2012 [29] to extract the deep feature of each video frame. Such features are known more robust against the drift across different domains.

Considering the trade-off of computation and accuracy, in our experiment we select ResNet-50 as the feature extractor and use the last layer’s features that are also called avg-pool whose dimension is 2048 and receptive field covers the whole image. Such deep features have rich semantic information and can well make up for the shortcomings of traditional methods by calculating the cosine similarity of two frames for measuring the difference in semantic-level. The default threshold is set to 0.95, and it means that the image pair will be identified as adjacent frame if the similarity score of two images is lower than 0.95.

The formula of cosine similarity:

\[
similarity(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}
\]  

(6)
bounding boxes and the corresponding confidence scores of the object which help us determine if the target is out of the screen.

In order to find a suitable threshold to determine whether the target has disappeared, we have tested the tracker on 20 video sequences of our target scenarios and then calculate the average confidence value of the tracking target in the process of its departure as the threshold.

Figure 3(a) illustrates an example video sequence which shows the score changing when people are walking out of the screen. Figure 3(b) illustrates the statistics of the changing of the confidence score when the object is leaving from the view. We marked the first frame as the beginning and the last frame as the ending, then average the confidence of these frames to get the threshold of this video. The mean threshold of all 20 target videos is about 0.73. Therefore, when the confidence of the target is less than 0.73 in 5 consecutive frames, we consider the target to leave the view. With this strategy, there is no failure case in our test set which includes 10 positive samples and 10 negative samples.

3) VISUAL OBJECT OF INTEREST ASSOCIATING

In order to implement the association function, we applied an IOU based matching mechanism by calculating the relative position of two objects. Since all the objects / people / face are detected separately, we should build up the relationship of them before doing the compliance audit (eg. match the uniform and name tag to the corresponding staff). Pipeline of standard checking for dressing and scene arrangement are as follows:

- Identify the participant and assign the ID.
- Detect target object (uniform, name tag and company logo), then mark the object position frame by frame.
- Associate the uniform and name tag to the corresponding staff.

We need to know the relationships between each body, face, uniform and name tag. In this paper, we can associate them by calculating the IOU of two objects, but in this use case, the traditional IOU is not suitable for measuring the overlap if one object is much larger than another. As we can see in Figure 4(a), $S_1$ is the small object, while $S_2$ is the big one. $S_\cap$ is the intersection area, but we modify the formulate here as $\text{IOU} = \frac{S_\cap}{S_1}$.

![Figure 4](attachment:image4.png)

**Figure 4.** Establish the association for all the objects/person in the screen (There are two bodies, two faces, two uniforms, and one tag were detected in the video frame. IOU of body1 and face1 is 1.0, then we bind them as id1. IOU of tag and body1 is 1.0 and that of uniform1 and body1 is 0.992, we determine that tag and uniform1 belongs to id1. IOU of body2 and face1 is 1.0, then we bind them as id2. IOU of uniform1 and body2 is 0.877, we determine that uniform2 belongs to id2).

In the three cases of matching, a certain face and body are considered to attach to the same person only if the IOU is over 0.9; a certain uniform is thought to belong to a certain body only if the IOU is over 0.8; if a person wears a work badge / logo, IOU of the object and the body should be 1.

C. PERSON-DEPARTURE DETECTION

The Measures requires that neither the staff nor the customer leave the camera screen during video recording, and we...
formulate this problem as a motion detection task that can be solved by detection and tracking technology. In the previous section, we have introduced YOLOv3 and Siamese-RPN to solve the detection and tracking problem separately. According to the Measures, here we have designed the algorithm pipeline of person-departure detection as Figure 5. The detector and tracker are not guaranteed to be 100% correct, and there may be false positives, so we are going to count the prediction result of five consecutive frames and then give the final decision if all of these frames have the same state. This strategy ensures the final prediction of this module has almost no error.

**FIGURE 5.** Algorithm pipeline of person-departure detection. The detector here is YOLOv3 and the tracker is Siamese-RPN. In order to make the final result not affect much by the false positive prediction, we detected the prediction results for 5 consecutive frames and then give the final decision if all of these frames have the same state.

**D. FACE IMAGE QUALITY ASSESSMENT**

To exactly identify the identity of the participants in the whole process, the Measures requires the face of the staff and the customer should be clear enough in the video. In this section, we propose two metrics to measure the quality of the face image:

1) **FACE SEMANTIC FEATURE**

We know the MTCNN network contains two heads, boundary box regression and classification, in which the classification branch gives the confidence of the foreground (face). The larger the confidence value is, the more obvious the face pattern is. Based on this experience, we directly use this output of the network as one metric to measure the quality of the face in the video sequence.

2) **VISUAL FEATURE**

Visual features here mainly include face size, posture and face image blurriness. To make a quantitative analysis of the evaluation results, a comprehensive scoring rule was established in this paper (see TABLE 1(a)). In this paper, the quality score was set from 0 to 100. The higher the score is, the better the quality is.

Facial size is one of the important factors which directly affects the quality of a face. We compress the face image at different resolution (bilinear interpolation), from 60 to 120, then use the face recognition model to calculate the similarity of the template face and the compressed face. The performance degrades a lot when the size is smaller than 60*60. Blur Detection works by using Laplace operator to perform sliding window operation on the detected face area and calculate the variance which can reasonably assess how blurry a face is, the Laplace operator is as follows:

\[
\begin{bmatrix}
0 & 1 & 0 \\
1 & -4 & 1 \\
0 & 1 & 1
\end{bmatrix}
\]

According to our experiment, we summarize the scoring scheme of facial size, blurriness, face pose and face semantic feature. For the size, blurriness, pose and confidence score of MTCNN, we set up an almost linear mapping from the original metric to the final score. Comprehensive scoring rule was established in this paper (TABLE 1(b)-1(e) lists the details of how to map each feature to face quality score).

Concretely, as MTCNN has detected five facial landmarks, as illustrated in Figure 6, in this part we directly use relative position of these landmarks for facial pose estimation. Pitch, yaw and roll are the three dimensions of movement when an object moves through a medium which used for measuring...
TABLE 1. Face image quality assessment (a) The weight coefficients of each feature in final scoring function (B) The corresponding scoring table for each feature.

| Feature | Weight of scoring | Blurriness | Pose (pitch/yaw/roll) | Semantic feature |
|---------|-------------------|------------|------------------------|------------------|
| Size    | 0.2               | 0.15       | 0.2/0.2/0.15           | 0.1              |

(b) THE FACE QUALITY SCORING TABLE FOR BLURRINESS

| Blurriness | Score |
|------------|-------|
| >600       | 100   |
| 500-600    | 90    |
| 400-500    | 80    |
| 300-400    | 70    |
| <60        | 60    |
| 100-200    | 40    |
| <100       | 0     |

(c) THE FACE QUALITY SCORING TABLE FOR FACE SIZE

| Face size (pixel) | Score |
|-------------------|-------|
| >120              | 100   |
| 100-120           | 80    |
| 80-100            | 60    |
| 60-80             | 20    |
| <60               | 0     |

(d) THE FACE QUALITY SCORING TABLE FOR FACE POSE

| Pose of face | Score |
|--------------|-------|
| Pitch >0.9   | 100   |
| 0.8-0.9      | 85    |
| 0.7-0.8      | 60    |
| 0.5-0.7      | 40    |
| 0.3-0.5      | 20    |
| <0.3         | 0     |

| Yaw >0.9     | 100   |
| 0.8-0.9      | 85    |
| 0.7-0.8      | 60    |
| 0.5-0.7      | 40    |
| 0.3-0.5      | 20    |
| <0.3         | 0     |

| Roll <0.3    | 100   |
| 0.3-0.5      | 80    |
| 0.5-0.65     | 60    |
| 0.65-0.8     | 20    |
| >0.8         | 0     |

(e) THE FACE QUALITY SCORING TABLE FOR SEMANTIC FEATURE

| Semantic feature (classification probability) | Score |
|----------------------------------------------|-------|
| >0.999                                       | 100   |
| 0.998-0.999                                  | 90    |
| 0.996-0.998                                  | 80    |
| 0.994-0.996                                  | 70    |
| 0.990-0.994                                  | 60    |
| <0.990                                       | 40    |

the facial pose here. The formula is expressed as follows:

\[
pitch = \frac{h_1}{h_2} \quad (8)
\]

where \(h_1 = y_3 - \frac{y_1+y_2}{2}, \quad h_2 = y_3 - \frac{y_4+y_5}{2}\).

\[
yaw = \frac{w_1}{w_2} \quad (9)
\]

where \(w_1 = x_3 - x_1, \quad w_2 = x_3 - x_2\).

\[
roll = \frac{|y_2 - y_1|}{h_1} \quad (10)
\]

E. SIGNATURE ACTION POSITIONING

Action recognition task involves the identification of different actions from video clips (a sequence of 2D frames) where the action may or may not be performed throughout the entire duration of the video. In this section, we use 3D CNN [19] model for action recognition. This model extracts features from both the spatial and the temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. In our practice, we made below improvements for the handwriting action prediction:

- In this application scenario, most of the videos contains two persons in a scene at the same time, which means it is not rigorous to recognize the action through the whole frames, as well as the complex background. We firstly crop a ROI region by the bounding box provided by the body detector and tracker, then feed the cropped ROI sequence to the network.

- We sample every 8 frames and take 16 consecutive frames each time as the network input. Once the writing action is detected for three consecutive times, we mark it as the starting point of the signing process.

- Then if the signature action is not recognized in three consecutive times, we regard it as the ending point.

IV. EXPERIMENTS AND CASE STUDY

In this section, we perform a collection of experiment on video tampering/integrity detection, object detection, object tracking and action recognition tasks, because this paper involves a commercial project and it is required by the client that the overall performance information shall not be disclosed. Therefore here we only report the component-wise performance numbers to obey the commercial constraints.

As we know, train a model from scratch requires a large dataset and takes a lot of time for converging, all network backbones used in this paper were pre-trained on the ImageNet dataset and then applied to the specific tasks.
B. Liu et al.: VCA for Compliance Audit in Finance and Security Industry

frames in different videos, and each set has 500 pairs. As you can see in TABLE 2, the higher value means the frame pairs are more similar. The parameter of MSSIM is set as we use a sliding window to divide the image into blocks. The window size set to 11. Considering the influence of the window shape on the block, Gaussian weights is used to calculate the mean, variance, and covariance of each window, and the standard deviation for Gaussian kernel set to 1.5.

| Adjacent frame | Interval frame | Irrelevant frame |
|----------------|----------------|------------------|
| MSSIM value    | 0.962          | 0.665            | 0.524            |

Meanwhile, we attack a normal video by frame insertion and frame deleting to test if there are obvious patterns when these attacks occur. The comparison results are shown in Figure 9. We can observe that the changing of MSSIM value of the normal video is very smooth in temporal, always lower than 0.02. In other words, the difference of the adjacent frame is usually less than 0.1, so we set the threshold of the MSSIM as 0.1. When inserting the attack video segment into the raw video, we can see there are two obvious peaks appeared in the Figure 9(b), so the attack frames can be easily located by analyzing the MSSIM value of the peaks (much larger than 0.1). Figure 9(c) shows the score changing when a short segment is deleted; there is obvious changes in values, and the peak value is around 0.2 which is also much higher than the threshold.

2) MSSIM UNDER NON-CONSTANT ILLUMINATION

While the light is changing, luminance compensation is applied for suppressing the interference caused by light changes. In the previous experiments, we only test the method with constant light condition, but in the realistic scenario, there are many reasons which cause the light condition change or drastic change, so only using MSSIM will cause false positive detection (light change will cause the value of MSSIM change a lot). Therefore, in this paper, we consider using a cascade detector with luminance compensation and semantic feature matching for accurate discrimination of false positive.

The structure of the cascade detector is shown as Figure 10. When the MSSIM value suddenly changes a lot, it will be considered as a potential tampering frame. Then we will compensate luminance for the neighbor frame of candidate frame (plus/minus the difference between the average value of these two frames) and calculate the MSSIM value again to do future decision.

In the processing of brightness compensation, in order to restore the intuitive visual information and avoid unnecessary interference, we perform a gamma correction on the darker frame. We dynamically select the value that makes the average brightness of the two frames closest to each other as the adjustment parameter (picked from 0.1, 0.2, ..., 0.9).
3) DEEP SEMANTIC FEATURE

As we mentioned before that the MSSIM based method may cause false alarms in some situations. The deep semantic feature extracted by ResNet-50 network, illustrated in Figure 2, is more focused on the high-level semantic information and insensitive to image noise. From Figure 11, we know it is beneficial to the traditional MSSIM based method. There are totally four peaks in the video sequence: p1 and p2 are brought by the luminance changing (false alarms), while p3 and p4 are brought by frame insertion. We know from Figure 11(a) that p1 and p2 were successfully suppressed by luminance compensation but the value is larger than the default MSSIM threshold (0.1), only using MSSIM criteria we cannot handle this situation. The bottom figure of Figure 11(a) shows the corresponding score from semantic feature similarity. We can easily identify the p1 and p2 are false alarms whose scores are larger than the default threshold of deep semantic feature, and also the attack frames p3 and p4. In Figure 11(b), all the peaks were brought by intense light changing. Although the luminance compensation module almost failed here, the deep semantic feature can successfully identify they are normal frames.

B. OBJECT DETECTION

In this part, we introduce the implementation details of the deep learning model training, include the datasets, the training hyper parameter selection and the corresponding evaluation results. The networks are trained using stochastic gradient descent (SGD) on GTX 2080Ti GPU.

1) CONSTRUCTING COMPLIANCE AUDIT OBJECT DATASET

As mentioned in previous section, we built the compliance audit object dataset by crawling images using image search engines, we manually defined the relevant key words for this scenario and then applied the crawling process. Since there are many noisy images in the raw collection, we spent much efforts to clean the data, the primary objective was to develop a robust dataset for optimizing the deep-learning based detection model. Our dataset features four categories representing generic types of the compliance audit target object, which we labelled for a total of 10.5K images and about 20.1K object bounding boxes, including person, uniform, name tag and company logo. The training set includes 8432 images, while the validation set includes 2108 images (The split of the original dataset is 80% training, 20% validation). The distribution of the dataset and corresponding performance of each category in this paper is listed at TABLE 3.

2) TRAINING THE MODEL

Data augmentation in convolutional neural network can robustly classify/detect objects even if they are placed in
different orientations as it has a property called invariance. More specifically, a CNN can be invariant to translation, color jittering, viewpoint, size or lighting noise (or a combination of the above). In our paper three types of data augmentation methods were used: affine transformation, gamma correction, and noise injection which make the model more robust to illumination changing, camera jitter and small motion in the scene. We also perform mean subtraction for the input of all models.

All the input image is resized to 608 × 608. We use the darknet-53 backbone pre-trained from ImageNet and fine-tune on our dataset. To determine the priors, YOLOv3 applies k-means cluster to generate the anchors. For our dataset, we follow the configuration of COCO, so the width and height of the pre-select anchors are (10 × 13), (16 × 30), (33 × 23), (30 × 61), (62 × 45), (59 × 119), (116 × 90), (156 × 198), (373 × 326). For the training parameters, we use a weight decay of 0.0001 and a momentum of 0.9. There are totally 150 epochs performed and the learning rate is 0.005 for the first 100 epochs, 0.001 at epoch 100-120 and 0.0002 for the next 30 epochs.

### C. HUMAN TRACKING

In this paper, the experiment is performed on VOT-2015 challenging tracking dataset [31]. The VOT2015 dataset consists of 60 sequences. The performance is evaluated in terms of accuracy (average overlap while tracking success-fully) and robustness (failure times). The overall performance is evaluated using Expected Average Overlap (EAO) which takes account of both accuracy and robustness. Besides, the speed is evaluated with a normalized speed (EFO). We didn’t report the EFO in our experiment, because this metric is very much related to the configuration of the hardware.

We use a modified AlexNet pre-trained from ImageNet with the parameters of the first three convolution layers fixed and only fine-tune the last two convolution layers in Siamese-RPN. The weight decay is 0.0001 and the momentum is 0.9. There are totally 80 epochs performed and the learning rate is decreased in log space from to. The final performance is a little better than the original paper (see TABLE 4), and this could be due to data augmentation and more training time.

### TABLE 4. The performance of Siamese-RPN on VOT-2015.

| EAO  | Accuracy | Failure |
|------|----------|---------|
| Siamese-RPN | 0.358  | 0.580  | 0.930  |
| Ours  | 0.365  | 0.587  | 0.920  |

### D. SIGNATURE ACTION POSITIONING

In this experiment, Kinetic-400 [32] was used for model training. Kinetic-400 contains 400 categories and the video segment is about 10 seconds. Each category contains about 400 to 1,150 videos where the writing category includes 735 segments. We spatially resize the sample at 112 × 112. The size of each sample is 3 × 16 × 112 × 112 (channels, frames, width, height), and each sample is horizontally flipped with 50% probability. All generated samples retain the same class labels as their original videos. Since we didn’t train the networks from scratch, the training parameters include a weight decay of 1e-5 and 0.9 for momentum. The initial learning rate is 0.001 and divide it by 10 after the validation loss saturates.

In this experiment, we used mobile phone to capture 50 video sequences, which totally include 50 writing clips, the length is about 10 seconds and the FPS is 25. The signature process of each video is about 3 seconds. At the inference stage, we sampled every 8 frames and take 16 consecutive frames each time as the network input.

As illustrated in Figure 8, we used the action recognition model to locate the ‘begin’ and ‘end’ of the writing action, here if the model’s prediction is ‘writing’ for 3 times in a row, we believe it’s the ‘begin’ timestamp, as well if the prediction is ‘non-writing’ for 3 times in a row, we think it’s the ‘end’ timestamp. Finally the IOU of the ground-truth and the prediction time window of writing action is calculated, if the IOU is larger than 0.5, this should be a good prediction. The average precision is 74% at IOU=0.5.

### V. CONCLUSION AND FUTURE WORK

The surge of deep learning over the last decades is impressive due to the strides it has enabled in the field of computer vision. In this work, we have introduced a deep learning-based video content analysis framework in compliance audit of stock brokerage to reduce the workload for the auditors and also speed up the audit process. The proposed framework integrates several state-of-the-art deep learning models in order to make the system stable and efficient. Although the system has not been fully deployed to large-scale business scenario, the prototype so far has proven its potential in real-world commercial applications.

For future work, to deal with the fundamental challenge for relatively small size training dataset problem and the varying drift of the objects for analysis in real-world visual audit setting, we plan to resort to image matching and registration [33], [34], as well as structure matching based approaches [34]–[36].

### REFERENCES

[1] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, “Handwritten digit recognition with a back-propagation network,” in Proc. Adv. Neural Inf. Process. Syst., 1990, pp. 396–404.

[2] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in Proc. 7th Int. Conf. Document Anal. Recognit., 2003, pp. 1–6.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[4] G. E. Hinton, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[5] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 153–160.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[9] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst. 2015, pp. 91–99.

[11] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[13] B. Li, J. Yan, W. Wu, Z. Zha, and X. Hu, “High performance visual tracking with siamese region proposal network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8971–8980.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Jun. 2010, pp. 1975–1981.

[16] T. Bao, S. Karmoshi, C. Ding, and M. Zhu, “Abnormal event detection and localization in crowded scenes based on PCANet,” Multimedia Tools Appl., vol. 76, no. 22, pp. 23213–23224, Nov. 2017.

[17] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan, and H. Snoussi, “AED-Net: An abnormal event detection network.” 2019, arXiv:1903.11891. [Online]. Available: http://arxiv.org/abs/1903.11891

[18] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 27th Asilomar Conf. Signals, Syst. Comput., Nov. 2003, pp. 1398–1402.

[19] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.

[20] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” ACM Comput. Surv., vol. 43, no. 4, pp. 1–43, 2011.

[21] J. C. V. Gemert, M. Jain, E. Gati, and C. G. M. Snoek, “APT: Action localization proposals from dense trajectories,” in Proc. Brit. Mach. Vis. Conf., 2015, p. 4.

[22] G. T. Papadopoulos, A. Xanconoupolos, and P. Daras, “Real-time skeleton-tracking-based human action recognition using kinect data,” in Proc. Int. Conf. Multimedia Modeling Dublin, Ireland: Springer, 2014, pp. 473–483.

[23] S. N. Paul and Y. J. Singh, “Survey on video analysis of human walking motion,” Int. J. Signal Process., Image Process. Pattern Recognit., vol. 7, no. 3, pp. 99–122, Jun. 2014.

[24] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 568–576.

[25] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4597–4605.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

BIN LIU received the bachelor’s degree in electronics engineering from the SWUST, China, in 2008, and the M.Sc. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. He has wide-ranging expertise in machine learning, computer vision, and computing in mathematics. His current research interests include the use of deep learning methodology for modeling, image classification, and object tracking.

MINGYU WU received the bachelor’s degree in electronics engineering from Sun Yat-sen University, Guangzhou, China, in 2018. He is currently pursuing the master’s degree in electrical engineering with Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision and machine learning, especially deep learning for image classification and object detection.

MINZEO TAO received the master’s degree in software engineering from Nanjing University. He is currently the Director of artificial intelligence with E- Capital Transfer Company Ltd. His research spans many aspects of programming language implementation, software development, high-performance computing, natural language processing, computer vision, and machine learning. In particular, his current research focuses on natural language processing, computer vision, and deep learning. He has recently been working on an intelligence robot application development and played a central role in the development of a conversational AI framework for building contextual assistants.
QIN WANG received the bachelor’s degree from Hunan University, in 2004. He is currently the Assistant General Manager and the Technology Director of the Business Unit of Intelligent Technology, E-Capital Transfer Company Ltd. His research interests include software engineering and machine learning, especially with applications in computer vision and natural language processing. He once worked as the key developer in both finance and the Internet industry.

LUYE HE received the bachelor’s degree in mathematics from Zhejiang University, in 2002. He is currently the General Manager of the Business Unit of Intelligent Technology, E-Capital Transfer Company Ltd. His research interests include software engineering and artificial intelligence, especially with applications in speech recognition and natural language processing. He once worked as the Vice General Manager of the Department of Information Technology, Shanghai Rural Commercial Bank (SRCB).

GUOLIANG SHEN is the Director of the Department of Technology and a Senior Engineer of automatic control with Zhejiang Zheneng Natural Gas Operation Company Ltd. He mainly engaged in the research and application of automation-related advanced algorithms, intelligent control, and model identification.

KAI CHEN (Member, IEEE) received the bachelor’s degree from the Huazhong University of Science and Technology, the master’s degree from Xi’an Jiao Tong University, and the Ph.D. degree from Shanghai Jiao Tong University. He is currently an Assistant Professor with the Department of Electrical Engineering, Shanghai Jiao Tong University. His research interests include data mining and computer vision, especially optical character recognition and its applications in the industry.

JUNCHI YAN (Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering and the Institute of Artificial Intelligence, Shanghai Jiao Tong University. Before that, he was a Senior Research Staff Member of IBM Research, China, where he started his career in April 2011, and once an Adjunct Professor with the School of Data Science, Fudan University. His research interests include machine learning and computer vision. He has served as Area Chair for ICPR 2020, and a Senior PC Member of CIKM in 2019. He will serve as the Area Chair of the CVPR, in 2021. He has published 60+ peer-reviewed articles in top venues and holds 20+ authorized U.S. patents. He has once been with IBM Watson Research Center, Japan NII, and Tencent/JD AI Lab, as a Visiting Researcher. He won the Distinguished Young Scientist of Scientific Chinese and the CCF Outstanding Doctoral Thesis. He serves as an Associate Editor for IEEE ACCESS, a Guest Editor (managing) for the IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEMS, Pattern Recognition Letters, and Pattern Recognition.