Text Co-Detection in Multi-View Scene

Chuan Wang, Huazhu Fu, *Senior Member, IEEE*, Liang Yang, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Multi-view scene analysis has been widely explored in computer vision, including numerous practical applications. The texts in multi-view scenes are often detected by following the existing text detection method in a single image, which however ignores the multi-view corresponding constraint. The multi-view correspondences may contain structure, location information and assist difficulties induced by factors like occlusion and perspective distortion, which are deficient in the single image scene. In this paper, we address the corresponding text detection task and propose a novel text co-detection method to identify the co-occurring texts among multi-view scene images with compositions of detection and correspondence under large environmental variations. In our text co-detection method, the visual and geometrical correspondences are designed to explore texts holding high pairwise representation similarity and guide the exploitation of texts with geometrical correspondences, simultaneously. To guarantee the pairwise consistency among multiple images, we additionally incorporate the cycle consistency constraint, which guarantees alignments of text correspondences in the image set. Finally, text correspondence is represented by a permutation matrix and solved via positive semidefinite and low-rank constraints. Moreover, we also collect a new text co-detection dataset consisting of multi-view image groups obtained from the same scene with different photographing conditions. The experiments show that our text co-detection obtains satisfactory performance and outperforms the related state-of-the-art text detection methods.**

***Index Terms*— Text co-detection, cycle consistency, epipolar geometrical guidance.**

I. Introduction

THE vision understanding in the multi-view scene has attracted considerable attentions in numerous tasks, like object recognition [1]–[3], categorization [4], modification [5], reconstruction [6], [7], retrieval [8], person and vehicle re-identification [9]–[11] and shape estimation [12], [13], which aims to explore complete representations from diverse facets. Compared to the single view scene, the multi-view scene not only holds a wealth of descriptions about the scene and objects in the scene, but also provides relationships of objects among multiple images. These relationships may include sequential relations for time series data, geometrical relations especially for rigidity objects, semantic relations for representative points of objects, and etc., which are absent in the single image scenes as a part of definitive information for object and scene understanding. For example, in the object discovering, the object-level correspondences not only enhance objects holding high probabilities, but also highlight objects that are not obvious in some images but distinct in other images. Correspondences describing relationships of points in objects also have the ability to better perform interior information of objects, e.g., the structure, due to the abundant and supplemental descriptions from multiple images holding diverse photographing views about objects. Therefore, exploring and incorporating correspondences with visual information shows a novel and attractive research direction for comprehensively scene understanding.

In the natural scenes, the text detection task has seen a surge of interests [14]–[20]. It has a wide range of application scenarios from understanding texts to the localization with text cues in the image-based or video-based scenes. Lots of works focus on text detection from the single image based on low-level manually-designed properties [18], [21]–[27], like Maximally Stable Extremal Regions (MSER) [28] and Stroke Width Transform (SWT) [18], or high-level classifiers [29]–[32]. These methods discover the discriminative text-related regions and then construct text candidates from them with a heuristic strategy. However, most of these researches have focused on the single image and may inevitably suffer limitations caused by restricted scene view and complex environmental factors. Multiple views of the scene have the ability of providing more visual textures and various facets of texts, which can be fully utilized to assist the text detection and understanding. Although there exist some text detection tasks for multi-view scenes, they are also often implemented by using the existing text detection method in individual image, which ignores the multi-view scenarios and corresponding relationships. To the best of our knowledge, our method is the first method for simultaneous detection of texts and their correspondences in multi-view scenes.

In this paper, the text co-detection task could be defined as, given a set of images consisting of multi-view images from the same scene, discovering text regions and identifying

Manuscript received March 29, 2019; revised November 5, 2019 and January 23, 2020; accepted January 29, 2020. Date of current version February 26, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0803701, in part by the National Natural Science Foundation of China under Grant 61733007 and Grant U1636214, in part by the Open Research Fund from the Shenzhen Research Institute of Big Data under Grant 2019ORF01010, in part by the Zhejiang Lab under Grant 2019NB0AB01, and in part by the Peng Cheng Laboratory Project of Guangdong Province under Grant PCL2018KP004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M. Ascenso. *(Corresponding author: Xiaochun Cao.)*

Chuan Wang and Xiaochun Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China, also with the Peng Cheng Laboratory, Cyberspace Security Research Center, Shenzhen 518055, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangchuan@iie.ac.cn; caoxiaochun@iie.ac.cn).

Huazhu Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi 300401, UAE (e-mail: hzfu@ieee.org).

Liang Yang is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China (e-mail: yangliang@vip.qq.com).

Digital Object Identifier 10.1109/TIP.2020.2973511

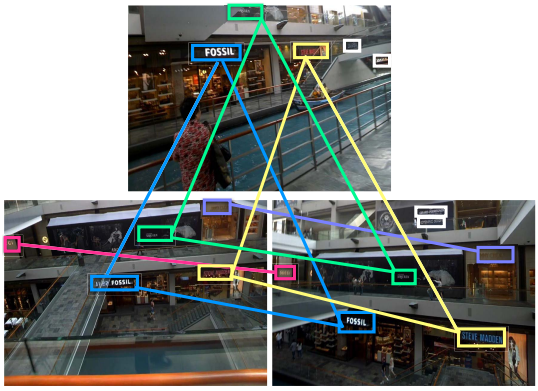


Fig. 1. The illustration of the multi-view text co-detection. Given a set of images photographed from the same scene with large perspective distortion, the goal of text co-detection is to discover text regions in images under the assistance from other relative images, and explore the corresponding relations of co-occurring texts. Texts linked with the same colored line are explored corresponding regions, while the individual texts are labeled by white color.

their relationships. In spite of the multi-view scenario, this task additionally requires building the corresponding relationships of text regions with a potentially large number of environmental distractors. It encourages us to detect text regions with the incorporation of visual information from the intra images and text correspondences from inter images. As shown in Fig. 1, by contrast to text detection in the single image, the text co-detection formulates correspondences among the text candidates from different images and identify the co-occurring texts in multi-view scene. In the practice, there have many vision applications providing opportunities to conduct text co-detection and further improve the performance. For example, in video analysis application, text co-detection has the ability of providing long-term correlations along the video frames by exploring the correspondences of co-texts. Since texts are common components in natural scenes and text co-detection can deal with views having large variation. Besides, considering the navigation task of the mobile robot and fully driverless task, both of which capture surrounding environments with multiple cameras covering different views, text co-detection can assist the understanding of surroundings from different views. What's more, since texts has no non-rigid deformations, exploring the correspondences of texts has the ability of providing landmark localization during multi-view 3D reconstruction.

In this paper, we propose a novel text co-detection method to simultaneously explore texts and their correspondences among multi-view images, with the multi-view geometry-correspondence guidance and multi-view cycle consistency. First, an instance-level similarity matrix based on visual features is generated to construct the consistency of the desired text correspondences. Besides, since visual similarities measure the images under similar situations like viewpoint, illumination, and perspective distortion, which are uncontrollable for texts, they may heavily result in high dissimilarities for related texts. We additionally construct the point-level epipolar geometrical relation, which is based on the observation that the multi-view images are obtained from the same scene and response the same geometry plane, as the guidance to conduct the exploitation of candidate correspondences. Incorporating visual feature similarities and geometrical guidance, we enforce the corresponding text candidate sets, each of which denotes candidate correspondence between all pairs of images, to satisfy the cycle consistency. Finally, the proposed framework could be formulated to a low-rank matrix recovery problem and solved effectively via binary semidefinite programming. In summary, we state the main contributions of this paper as follows:

- We address a new task, text co-detection, to discovery the co-occurring texts from the multi-view scene images.
- A novel text co-detection framework is proposed, which explores the visual similarity, geometrical information, and cycle-consistent candidate correspondences.
- The framework is formulated as a low-rank matrix recovery problem and solved with binary semidefinite programming.
- For the evaluation of text co-detection, a new multi-view dataset with the ground truth is collected, which contains 240 image sets and 727 images in total.
- We propose four evaluation metrics related to text co-detection task to present the performance evaluation.
- Our method achieves superior performance and not only includes co-occurring texts of multi-view images, but also has the ability to present correspondences for each co-occurring text.

II. Related Work

A. Text Detection Problem

Under the natural scene, texts always have a full of challenges suffering from both intrinsic difficulties of texts and extrinsic obstacles of the natural environment. In the intrinsic aspect, texts have different visual information, such as nonuniform stroke widths, different colors, various font styles, and changeable scales. These factors may bring big trouble in capturing consistent and typical representations of texts. In the extrinsic aspect, they always perform extremely sensitive to external factors, which may act big interference on obtaining representative and uniform features. In detail, scene texts always are very sensitive to uncontrolled environments and confusing background. They are also subtle to various conditions like scale, orientation, perspective distortion, low resolution, and non-uniform illumination. All of these factors may result in texts with diverse appearances and therefore give rise to failures for scene text detection.

To overcome difficulties and obtain good text detection results, a lots of works are proposed ranging from character-based to region-based detection strategies, which often consider relationships in text characters, among text characters or in text regions. For example, Jung *et al.* analyze the visual [14] and Epshtein *et al.* [18] focus on explore effective stroke filters, which consider the relationships of strokes like gray consistency and symmetry characteristic. Yin *et al.* [26] analyze the visual feature similarity between character candidates to construct text candidates. Zhu and Zanibbi [33] consider relationships of characters like color and size uniformity and

distance distribution to construct text regions. Gao *et al.* [15] focus on texts with the ordered sequence, which contains the ordered relationship with similar properties. Tian *et al.* [34] also incorporate ordered sequential relationships, i.e., regions within a text line share similar visual information, to improve text localization accuracy. Shi *et al.* [29] consider the relationship between neighboring text regions and formulate relationships by linkages of neighboring regions. They both employ the local linkage relation on a smaller scale and the high-level global linkage relation on a larger scale. Liao *et al.* [35] propose to directly predict the quadrilateral bounding boxes for arbitrary-oriented scene text detection and utilize "long" convolutional kernels to handle long text lines. Lyu *et al.* [36] deal with arbitrary-shape text spotting with an end-to-end trainable neural networks. They propose a four-components framework to classify and segment text proposals. They consider both global word map and character maps to provide accurate localization. Liao *et al.* [37] propose to use two network branches of different designs separately dealing with text classification and regression. They design the regression branch to extract rotation-sensitive features and the classification branch to extract rotation-invariant features. Long *et al.* [38] propose a flexible text representation, named TextSnake, dealing with texts in horizontal, oriented and curved forms. They describe a text with a sequence of ordered, overlapping disks centered at symmetric axes, which is defined by radius and orientation. Since each image has its corresponding radius and orientation maps, they conduct the detection as a segmentation task to obtain instance-level segmentation.

However, both local linkage based and region-based relationship based methods set up the conduction based on the prior knowledge that texts in the scene are always continuous and have slight variations. When they encounter large perspective variations and distortions, they may not catch correct relationships and fail to explore exact text locations. In addition, since texts have diverse appearances and are sensitive to environmental conditions, simply considering the visual feature similarities between pairs of texts may fail to discover corresponding ones. Above all, seeking to explore multi-way information in scenes with large variation has high significance on the text detection task. In this paper, we provide a novel application aspect for the text detection, i.e., multi-view text co-detection, which focuses on handling the co-occurring text detection under images with large perspective differences and environmental changes.

B. Object Co-Segmentation, Co-Detection and Joint Matching

Recently, simultaneously exploring multiple images have been proposed to consider broader visual information for various tasks, e.g., co-segmentation [39]–[43], co-saliency detection [44], [45], object co-detection [46], [47] and joint object matching [48], [49]. The "co-method" aims to jointly exploit multiple instances of a target from a set of images, most of which are obvious and prominent in images. It leverages appearance characteristics of instances in multiple images and is provenly key to improve performance. The joint object matching task aims to simultaneously estimate maps among a collection of objects and has become an emerging field.

Fu *et al.* [40] focus on co-segment multiple foreground objects from videos, under the consideration of the intra-video coherence and inter-video consistency of foregrounds. Wang *et al.* [41] design to co-segment foregrounds from a collection of images with the strategy that coherent foregrounds could construct a tight clique. On the object co-detection task, Bao *et al.* [50] incorporate a unified objective function on both detection and matching, and consider object- and part-level correspondences in pair of images. However, they require a relatively large amount of labeled matching objects in training step. Shi *et al.* [51] develop a human co-detection and labeling framework in a semi-supervised learning manner. Hayder *et al.* [52] construct a fully-connected Conditional Random Field (CRF) in which nodes represent the candidate labels, and the edges encode the appearance similarity between two candidates. They encode the appearance similarity as a mixture of Gaussian kernels and require supervised learning step to estimate the weights of kernels. However, in practice providing manually annotated co-occurrences (correspondences) are limited and seldom offered. Besides, most of them only consider corresponding or not-corresponding relationships between two candidates from two images, and relax the requirement that one candidate in one image only matches one candidate in the other image if there exists correspondence. On the joint matching task, the feature matching [53] or the pixel-wise flow computation [54] employ the cycle consistency as an additional constraint, which aims to identify incorrect matches from bad cycles, but having difficulties on seeking global solutions for feature-point-wise or pixel-wise graph relations.

However, compared to object co-segmentation, co-detection and joint matching, scene text co-detection task meets more difficulties. Firstly, texts have substantial diversity from generic objects, which do not hold well-defined enclosed boundary and center. Besides, scene text is diverse and may perform entirely different fonts, scales, and orientations. Further, scene text always accompanies with the very complex environment. For example, cluttered textures like signs, fences, and bricks are confused as text. Thus, texts are virtually indistinguishable from background textures. Therefore, the proposed text co-detection task requires elaborated operations to handle abundant and variant visual and view information of multi-view scenes.

To deal with the special and difficult text co-detection task, in this paper we design a novel text co-detection method based to simultaneously explore texts and their correspondences. Compared with previous object co-detection algorithms [50]–[52], the proposed co-detection method requires no training step with groundtruth text locations and correspondences. The proposed method firstly considers the one-to-one constraint since a text region in one image only have one corresponding text in the other image if the correspondence exists. Besides the pairwise relationships, the proposed method additionally incorporates cycle consistency in co-detection task to ensure the propagation of correspondences among multiple images. Further, the proposed text co-detection method specifically incorporate scene projection relationship to provide

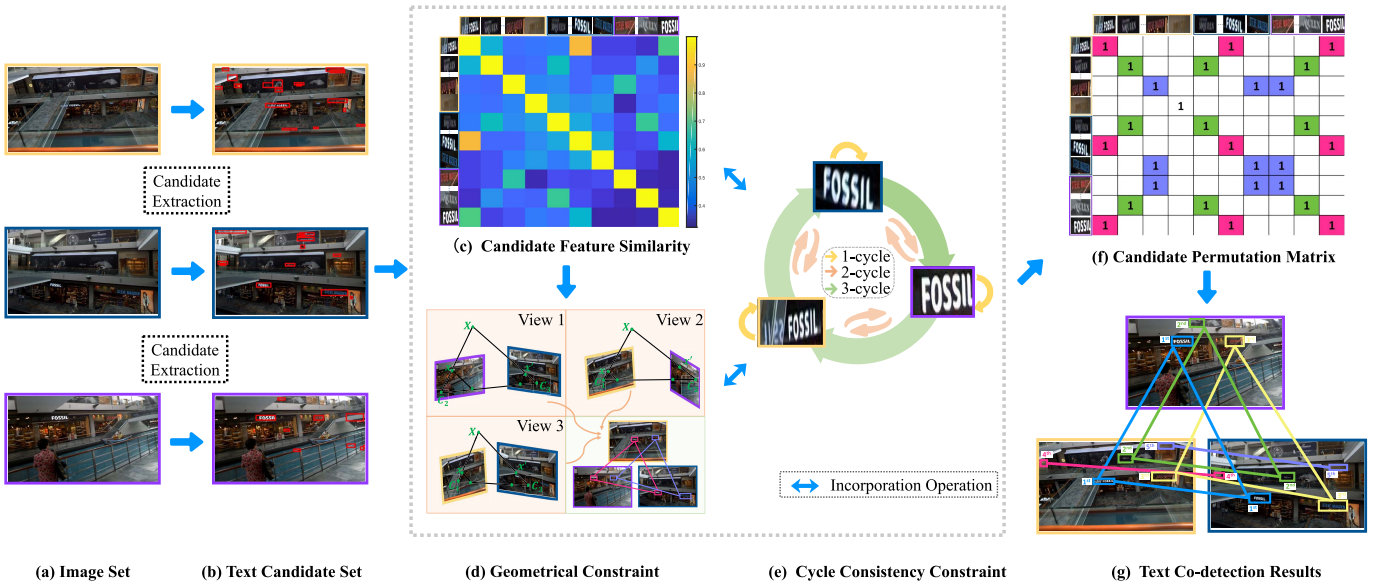


Fig. 2. Illustration of the proposed text co-detection method. Given a set of multi-view images (a) with their text candidates (b), we firstly compute the similarities of each candidate pair based on visual feature representations (c). We utilize the epipolar geometrical information based on points within texts as additional guidance (d). Then the cycle consistency among candidates (e) is incorporated to preserve the correspondences among multi-view images. We jointly explore corresponding candidates under the permutation matrix formulation (f) and identify the corresponding texts (g) (candidates linked by the line with the same color).

much guidance since texts have no nonrigid deformation. All these aspects differ from the characteristics of object co-detection.

III. Text Co-Detection Formulation

We start by introducing the text co-detection definition in this section, and subsequently give a co-detection framework integrating the visual feature similarity, the geometrical guidance and the cycle consistency for multi-view images.

A. Problem Statement and Overview

The text co-detection task is designed to simultaneously explore the co-occurring texts and their correspondences among multi-view images. Texts that are presented among images with different views are treated as the desired co-occurring texts. As shown in Fig. 2 (a), texts "FOSSIL", "STEVE MADDEN" and "MQUEEN" in three images from the same scene are regarded as the co-occurring texts.

We propose to explore the co-occurring texts via discovering corresponding texts among multi-view images. As shown in Fig. 2, given N multi-view images $\{I_1, I_2, \ldots, I_N\}$ and text candidate sets $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$, the pairwise candidate visual feature similarity $\mathcal{S}_{ij}^v$ represents the corresponding probabilities between candidates from different views (I_i, I_j). Besides, we employ geometrical relation f^g, which is produced by the epipolar geometry, on multi-view images for the geometrical consistency. With visual and geometrical consistencies, we seek a correspondence matrix $\mathbf{X}$ for text candidates, which indicates relations for all pairwise candidates of the images. It is formulated as the arrangement of pairwise image correspondence $\mathbf{X}_{ij} \in \{0,1\}^{q_i \times q_j}$. Each element x_{ij}^{mn} in $\mathbf{X}_{ij}$ indicates whether two candidates c_i^m and c_j^n, which respectively come from I_i and I_j, have the relationship that assigns x_{ij}^{mn} to 1. The resulting co-occurring text candidates, as presented via the linked candidates in Figure 2 (g), not only have high pairwise visual affinities and geometrical scores but also are forced to be consistency among multiple images. To make the indication to be clear, we summarize the notations of variables in the paper in Table I.

TABLE I
Notations

Symbol	Description
N	the number of images
Q	the number of candidates of all images
I_i	the i^{th} image in the multi-view image set
$\mathcal{C}_i$	the text candidate set of the i^{th} image
q_i	the number of candidates in $\mathcal{C}_i$
c_i^m	the m^{th} candidate for image I_i
f^v	the geometrical consistency for all pairs of images
$\mathbf{d}_i$	the indication of feature points for the image I_i
$\mathbf{D}_i$	the permutation matrix of indication $\mathbf{d}_i$ by $\mathbf{D}_i = [\mathbf{d}_i \mathbf{d}_i \mathbf{d}_i]$
$\mathbf{P}^i$	the geometrical feature points of the image I_i
$\mathbf{M}_i$	the indication matrix for candidates to feature points in image I_i
$\mathbf{F}_{ij}$	the fundamental matrix of epipolar geometry between image I_i and I_j
$\mathbf{S}^v$	the visual feature similarity for all images
$\mathbf{S}_{ij}^v$	the visual feature similarity between image I_i and I_j
$\mathbf{X}$	correspondence for all images
$\mathbf{X}_{ij}$	image correspondence between image I_i and I_j
x_{ij}^{mn}	candidate correspondence between m^{th} candidate in image I_i and n^{th} candidate in image I_j

B. Co-Detection With Text Candidates

Given N images from multiple view scenes, we design to explore co-occurring texts with the assistance of correspondences of text candidates.

We first build the text candidate set $\mathcal{C}_i$ for each image I_i. One general assumption is that the candidate set $\mathcal{C}_i$ could cover most of the text regions in the image I_i. In this paper, the text candidates are generated by integrating the deep-learning representations and text-specifically hand-crafted features. The Single Shot Text Detection (SSTD) [30] network is utilized to explore the deep non-linear text representations. The threshold that control candidate selection and filtering in SSTD is relaxed from 0.6 to 0.3 to cover as more as possible candidates with high detection recall. SWT interesting points [18] is used to obtain text-specific regions, group neighboring SWT points and subsequently filter these regions by discarding regions whose width or height is less than 10 pixels. The remaining regions from deep and text-specific representations are collected to build the candidate set $\mathcal{C}_i$ with q_i candidates.

We desire to find a matrix $\mathbf{X}_{ij}$, which is composited by the elements in candidate sets $\mathcal{C}_i$ and $\mathcal{C}_j$, to indicate the presence and the corresponding relationship of co-occurring texts. Since the correspondences between texts are the one-to-one relationship, we represent the matrix $\mathbf{X}_{ij}$ as a partial permutation matrix $\mathbf{X}_{ij} \in \{0,1\}^{q_i \times q_j}$ as:

$$\mathbf{X}_{ij}\mathbf{1} = \mathbf{1}, \quad \mathbf{X}_{ij}^T\mathbf{1} = \mathbf{1}, \tag{1}$$

where $\mathbf{1}$ is an all-one vector. In spite of the permutation matrix constraints, the correspondence $\mathbf{X}_{ij}$ satisfies the self-corresponding and symmetric constraints:

$$\begin{aligned} \mathbf{X}_{ii} &= \mathbf{I}_i, \quad 1 \le i \le N \\ \mathbf{X}_{ij} &= \mathbf{X}_{ji}^T, \quad 1 \le i < j \le N, \end{aligned} \tag{2}$$

where $\mathbf{I}_i$ denotes the identity matrix of image I_i. Besides, the correspondences $\mathbf{X}_{ij}$ is sparse since at most one value in each row of $\mathbf{X}_{ij}$ is valued.

To incorporate all image pairs, the correspondences of multiple images are stored in $\mathbf{X}$ and constructed as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1N} \\ \mathbf{X}_{12}^T & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2N} \\ \vdots & \vdots & \mathbf{X}_{(N-1),(N-1)} & \mathbf{X}_{(N-1),N} \\ \mathbf{X}_{1N}^T & \vdots & \mathbf{X}_{(N-1),N}^T & \mathbf{X}_{NN} \end{pmatrix}. \tag{3}$$

To discover the correspondences $\mathbf{X}$, we enforce the consistency between the correspondence $\mathbf{X}$ and the text visual feature similarity $\mathbf{S}^{\mathbf{v}}$, which is computed by the normalized correlation between text deep feature representations of each image pair. The correspondence $\mathbf{X}$ also subjects to the constraints in Eq.s 1, 2 and the sparsity as:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{argmin}} \ \|\mathbf{X} - \mathbf{S}^v\|_F^2 + \alpha\|\mathbf{X}\|_0, \\ &s.t., \ \mathbf{X}_{ij}\mathbf{1} = \mathbf{1}, \quad \mathbf{X}_{ij}^T\mathbf{1} = \mathbf{1}, \\ &\qquad \mathbf{X}_{ii} = \mathbf{I}_i, \quad 1 \le i \le N, \\ &\qquad \mathbf{X}_{ij} = \mathbf{X}_{ji}^T, \quad 1 \le i < j \le N, \end{aligned} \tag{4}$$

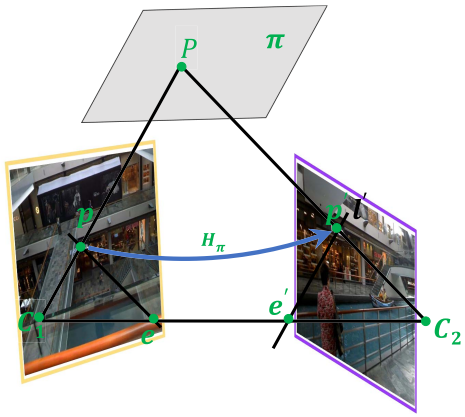


Fig. 3. Illustration of the epipolar geometry. The point $\mathbf{p}$ in one image is transferred through the plane π to the matching point $\mathbf{p}'$ in the other image. The epipolar line l' through $\mathbf{p}'$ is obtained by linking $\mathbf{p}'$ to the epipole e'. We can write $\mathbf{p}' = \mathbf{H}_\pi\mathbf{p}$ and $l' = [e'] \times \mathbf{p}' = [e'] \times \mathbf{H}_\pi\mathbf{p} = \mathbf{F}\mathbf{p}$ where $\mathbf{F} = [e'] \times \mathbf{H}_\pi$ is the fundamental matrix.

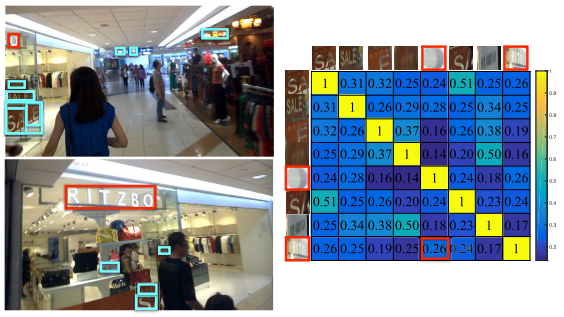
where α is the weight of sparsity of $\mathbf{X}$ and predefined with 0.5 in this paper.

C. Geometrical Consistency With Epipolar Geometry

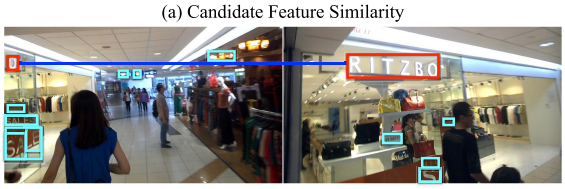
Given a pair of images captured from the same scene with different views, the intrinsic projective geometry between different views, i.e., the epipolar geometry, is usually motivated by considering the estimation of corresponding points between the two views. The property of epipolar geometry inspires us that feature points in corresponding text candidates have a high probability to be the corresponding points. Besides, partial presence of text candidates may induce low visual feature similarity and incorrect correspondence (as shown in Fig. 4), which require additional information to provide rectification about incorrect correspondences. Therefore, we introduce the feature point correspondence to enhance the candidate correspondence. In the following, we firstly give a brief introduction to epipolar geometry and its algebra representation. Then we illustrate the formulation of geometrical consistency employed in the paper.

The epipolar geometry is the intrinsic projective geometry between two different views, regardless of how the scene structure is changed. It describes that the corresponding points $\mathbf{p} = \{x, y, 1\} \in \mathbb{R}^{3\times1}$ and $\mathbf{p}' = \{x', y', 1\} \in \mathbb{R}^{3\times1}$ in images of two views are coplanar in the epipolar plane π with their space point P (the point in 3D) and the camera centers $\mathbf{C}$ and $\mathbf{C}'$, as shown in Fig 3. Considering that points $\mathbf{p}$ and $\mathbf{p}'$ are corresponding, the point $\mathbf{p}'$ must lie on the epipolar line $\mathbf{l}'$, which is defined by the intersection of π and the image plane of $\mathbf{p}'$ [55]. Thus, there is a map $\mathbf{p} \mapsto \mathbf{l}'$ from the point $\mathbf{p}$ in one image to its corresponding epipolar line in the other image. This projective mapping from points to lines is represented by a fundamental matrix $\mathbf{F}$ with $l' = \mathbf{F}p$. With the corresponding point p' lying on l' we have $0 = p'^T l' = p'^T\mathbf{F}p$, which is used as geometrical guidance to construct geometrical information and boost the correspondences.

Based on the defined geometrical guidance, we formulate the geometrical consistency by integrating the geometrical



(a) Candidate Feature Similarity



(b) Prior from Epipolar Geometry Constraint

Fig. 4. Partial presence of text candidates induces low visual feature similarity and incorrect correspondence, as shown in (a). The guidance obtained from the epipolar geometry has the ability to guide the correspondence, as shown in (b).

guidance and candidate correspondence. The fundamental matrices $\mathbf{F}_{ij}$'s for image pairs are calculated beforehand. For each image I_i, we extract the interest feature points with Scale Invariant Feature Transform (SIFT) [56] and SWT [18]. The SIFT and SWT feature points locating inside the candidates are preserved as feature points $\mathbf{P}_i$ for each image I_i. The similarities of pairwise points are represented by SIFT feature distance and used to find the initial matching relation between points. Then the Random Sample Consensus method (RANSAC) is applied to estimate fundamental matrices $\mathbf{F}_{ij}$'s for image pairs $(I_i, I_j), \forall i, j \in \{1, N\}, i \neq j$.

With the estimated fundamental matrices $\mathbf{F}_{ij}$'s, we calculate the geometrical consistency for image pairs in the set. Suppose we have the correspondence matrix $\mathbf{X}$ for the image set. The sub-matrix $\mathbf{X}_{ij} \in \mathbb{R}^{q_i \times q_j}$ for an image pair $(I_i, I_j), i, j \in \{1, N\}$ is the correspondences of the candidates in $\mathcal{C}_i$ and those in $\mathcal{C}_j$. We use feature points located in corresponding candidates as candidate points of each image, whose selection is indicated by $\mathbf{d}_i = \mathbf{M}_i \mathbf{X}_{ij} \mathbf{1}_j^T$. $\mathbf{M}_i \in \mathbb{R}^{p_i \times q_i}$ is the mapping matrix of feature points to text candidates with p_i is the number of feature points and q_i is that of text candidates of image I_i. $\mathbf{1}_j$ is an all-one vector whose length is q_j. We permutate the point selection d_i by $\mathbf{D}_i = [\mathbf{d}_i\ \mathbf{d}_i\ \mathbf{d}_i]$ to have the same size of feature points $\mathbf{P}_i$. Then the selected feature points $\hat{\mathbf{P}}_i$ for image I_i is represented as $\hat{\mathbf{P}}_i = \mathbf{P}_i \odot \mathbf{D}_i$. Thus, the geometrical consistency can be imposed on selected candidates by minimizing the following term:

$$f^g = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \| \hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j^T \|_F^2. \tag{5}$$

The minimization of Eq. 5 aims to ensure that feature points in the corresponding candidates have lower fundamental responses, i.e., the value of $\| \hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j^T \|_F^2$, than those in the unrelated candidates.

D. Co-Detection With Cycle Consistency

To handle a bunch of images, cycle consistency [57]–[59] is employed to ensure the correspondences among multiple images. The definition of the cycle consistency is, if a candidate c_i^m in image I_i has correspondences with candidates c_j^n in image I_j and c_k^t in image I_k, i.e., $x_{ij}^{mn} = x_{ik}^{mt} = 1$, there also exists a correspondence between candidates c_j^n and c_k^t in image I_j and I_k, i.e., $x_{jk}^{nt} = 1$.[1]

$$\mathbf{X}_{ij} = \mathbf{X}_{kj} \mathbf{X}_{ik}, \quad 1 \leq i < j < k \leq N. \tag{6}$$

The desired $\mathbf{X}$ holds low-rank and positive semidefinite properties. Low-rank property of $\mathbf{X}$ is straightforward due to the matrix expression of cycle consistency shown in Eq. 6. The positive semidefinite property is also straightforward as $\forall \mathbf{z} \in \mathbb{R}^{NQ \times 1}, \mathbf{z}^T \mathbf{X} \mathbf{z} = \mathbf{z}^T \mathbf{V}_i^T \mathbf{V}_i \mathbf{z} = \| \mathbf{V}_i \mathbf{z} \|_{\mathcal{F}}^2 \geq 0$. $\mathbf{V}_i$ is the matrix representing as $\mathbf{V}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iN})$.[2] The requirements of the solution of $\mathbf{X}$ are written as:

$$\mathbf{X} \succeq 0, \quad \text{rank}(\mathbf{X}) \leq r, \tag{7}$$

where r is the rank of $\mathbf{X}$ and assigned with the maximal number of candidates of an image. Therefore, the problem of exploring cycle-consistent correspondence matrix is simplified and equivalently constructed as solving a binary positive semidefinite matrix with the low-rank constraint. The diagonal blocks of $\mathbf{X}$ are identity matrices, and off-diagonal blocks are permutation matrices.

E. Formulation

Incorporating the geometrical guidance in Eq. 5 and the cycle consistency constraint into the objective function in Eq. 4, we can summarize that the desired correspondence $\mathbf{X}$ should satisfy requirements: 1) consistent with the visual feature similarity $\mathbf{S}^v$ and the geometrical guidance (Eq. 5); 2) sparse due to the permutation matrix formulation in Eq. 1; 3) low-rank and positive semidefinite (Eq. 7). Hence, the objective function is written as:

$$\begin{aligned} \underset{\mathbf{X}}{\text{argmin}} \ & \| \mathbf{X} - \mathbf{S}^v \|_F^2 + \alpha \| \mathbf{X} \|_0 + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \| \hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j \|_F^2 \\ & + \gamma \, \text{rank}(\mathbf{X}), \\ s.t., \ & \mathbf{X}_{ij} \mathbf{1} = \mathbf{1}, \ \mathbf{X}_{ij}^T \mathbf{1} = \mathbf{1}, \\ & \mathbf{X}_{ii} = \mathbf{I}_i, \quad 1 \leq i \leq N, \\ & \mathbf{X}_{ij} = \mathbf{X}_{ji}^T, \quad 1 \leq i < j \leq N, \\ & \mathbf{X} \succeq 0, \\ & \text{rank}(\mathbf{X}) \leq r, \end{aligned} \tag{8}$$

where γ controls the weight of the rank.

To assure that the cycle-consistency is actually convex, we relax the permutation matrix constraint in Eq. 1 of $\mathbf{X}$ with the doubly stochastic constraint. That is, each element of $\mathbf{X}_{ij}$

[1] As we compute the correspondences between all pairs of candidates, taking 1-cycles (the self-corresponding constraint), 2-cycles (the symmetric constraint) and 3-cycles (relationships among three candidates from three different images) is sufficient.

[2] The detail proof is shown in [60].

takes a real value between 0 and 1, and rows and columns of each block of $\mathbf{X}$ sum to 1.

$$\mathbf{0} \leq \mathbf{X}_{ij} \leq \mathbf{1}, \mathbf{0} \leq \mathbf{X}_{ij}^T \leq \mathbf{1}, \sum_i \mathbf{X}_{ij} = \sum_j \mathbf{X}_{ij} = 1. \quad (9)$$

Besides, to make the optimization tractable, we relax the constraint of the rank of $\mathbf{X}$ by replacing it with the nuclear norm of $\mathbf{X}$ (sum of singular values of $\mathbf{X}$). To preserve the sparsity of $\mathbf{X}$, we also relax $\ell 0$-Norm to $\ell 1$-Norm that equals to the sum of absolute values in $\mathbf{X}$. The objective function in Eq. 8 is rewritten as:

$$\begin{aligned} \underset{\mathbf{X}}{\operatorname{argmin}} \ & \|\mathbf{X} - \mathbf{S}^v\|_F^2 + \alpha \|\mathbf{X}\|_1 + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \|\hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j\|_F^2 \\ & + \gamma \|\mathbf{X}\|_*, \\ s.t. \ & \mathbf{X} \in \mathcal{A}, \end{aligned} \quad (10)$$

where $\mathcal{A}$ is the set of matrices including low-rank, positive semidefinite and doubly stochastic constraints Eq. 9.[3]

Since elements in $\mathbf{X}$ are no less than 0, thus the $\ell 1$-Norm of $\mathbf{X}$ is replaced by the inner product of an all-one vector $\mathbf{1}$ and $\mathbf{X}$. The objective function defined in Eq. 10 is converted as:

$$\begin{aligned} \underset{\mathbf{X}}{\operatorname{argmin}} \ & -2 < \mathbf{X}, \mathbf{S}^v > + \alpha < \mathbf{1}, \mathbf{X} > \\ & + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \|\hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j\|_F^2 + \gamma \|\mathbf{X}\|_* \\ & = \underset{\mathbf{X}}{\operatorname{argmin}} < \mathbf{W}, \mathbf{X} > \\ & + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \|\hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j\|_F^2 + \gamma \|\mathbf{X}\|_*, \\ s.t. \ & \mathbf{X} \in \mathcal{A}, \end{aligned} \quad (11)$$

where $< \cdot, \cdot >$ is the inner product. $\mathbf{W} = \alpha \mathbf{1} - 2\mathbf{S}^v$. By solving Eq. 11, the most repeatable text candidates in the image set will be obtained and matched with a consistent way.

F. Optimization

The nuclear norm minimization in Eq. 11 is convex and always be solved with the proximal method [62] or Alternating Direction Method of Multipliers (ADMM) [63], each of which is based on iterative singular value threshold [64]. Since using singular value decomposition in each iteration is expensive, we follow the previous works [61], [65] to solve the problem by replacing the variable $\mathbf{X}$ with two matrices $\mathbf{A}$ and $\mathbf{B}$ whose dimensions are smaller than $\mathbf{X}$. In the objective function we let $\mathbf{X} = \mathbf{A}\mathbf{B}^T$ and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{Q \times r}$. These two new variables has smaller dimension $r < Q$. Besides, we make the second replacement for $\mathbf{X}_{ij}$'s with $\mathbf{Y}_{ij}$'s in the term of geometrical consistency since each sub-problem in the block coordinate descent will be much easier to solve. The optimization problem in Eq. 11 is rewritten as:

$$\begin{aligned} \underset{\mathbf{X},\mathbf{Y}}{\operatorname{argmin}} \ & < \mathbf{W}, \mathbf{X} > + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \|\hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j\|_F^2 \\ & + \gamma \|\mathbf{A}\mathbf{B}^T\|_* + \frac{\rho}{2} \|\mathbf{Y} - \mathbf{X}\|^2, \\ s.t. \ & \mathbf{X} = \mathbf{A}\mathbf{B}^T, \ \mathbf{X} \in \mathcal{A} \\ & \mathbf{Y}_{ij} \in \mathbb{P}^{q_i \times q_j}, 1 \leq i \leq N, \end{aligned} \quad (12)$$

With the following equation [66],

$$\|\mathbf{X}\|_* = \min_{\mathbf{A},\mathbf{B}:\mathbf{A}\mathbf{B}^T = \mathbf{X}} \frac{1}{2} (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2). \quad (13)$$

The final objective function is formulated as:

$$\begin{aligned} \underset{\mathbf{X},\mathbf{Y}}{\operatorname{argmin}} \ & < \mathbf{W}, \mathbf{X} > + \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \|\hat{\mathbf{P}}_i \mathbf{F}_{ij} \hat{\mathbf{P}}_j\|_F^2 \\ & + \frac{\gamma}{2} \|\mathbf{A}\|_F^2 + \frac{\gamma}{2} \|\mathbf{B}\|_F^2 + \frac{\rho}{2} \|\mathbf{Y} - \mathbf{X}\|^2 \\ s.t. \ & \mathbf{X} = \mathbf{A}\mathbf{B}^T, \mathbf{X} \in \mathcal{A} \\ & \mathbf{Y}_{ij} \in \mathbb{P}^{q_i \times q_j}, 1 \leq i \leq N. \end{aligned} \quad (14)$$

To solve the optimization in Eq. 14, we alternately update $\mathbf{X}$, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Y}$ in the following manner. We firstly jointly compute the update for $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{B}$ with given $\mathbf{Y}$. Subsequently, we update the value of $\mathbf{Y}$ with given $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{B}$.

To update $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{B}$, Eq. 14 is rewritten as

$$\begin{aligned} \underset{\mathbf{X},\mathbf{A},\mathbf{B}}{\operatorname{argmin}} \ & < \mathbf{W}, \mathbf{X} > + \frac{\gamma}{2} \|\mathbf{A}\|_F^2 + \frac{\gamma}{2} \|\mathbf{B}\|_F^2 + \frac{\rho}{2} \|\mathbf{Y} - \mathbf{X}\|^2 \\ s.t. \ & \mathbf{X} = \mathbf{A}\mathbf{B}^T, \ \mathbf{X} \in \mathcal{A}, \end{aligned} \quad (15)$$

and ADMM [63] is applied to solve the Eq. 15. The augmented Lagrangian of Eq. 15 is:

$$\begin{aligned} \mathcal{L}_\mu(\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbf{Z}) = & < \mathbf{W}, \mathbf{X} > + \frac{\gamma}{2} \|\mathbf{A}\|_F^2 + \frac{\gamma}{2} \|\mathbf{B}\|_F^2 \\ & + < \mathbf{Z}, \mathbf{X} - \mathbf{A}\mathbf{B}^T > + \frac{\mu}{2} \|\mathbf{X} - \mathbf{A}\mathbf{B}^T\|_F^2 \\ & + \frac{\rho}{2} \|\mathbf{X} - \mathbf{Y}\|^2. \end{aligned} \quad (16)$$

$\mathbf{Z}$ is the Lagrange multiplier and μ is a parameter which controls the step size in optimization. The ADMM alternately updates each primal variable by minimizing $\mathcal{L}_\mu$. The dual variable is updated by gradient ascent with fixing all others.

The minimization of $\mathcal{L}_\mu$ for $\mathbf{A}$ and $\mathbf{B}$ turn to be a regularized least squares problem with a closed-form solution, with the following forms:

$$\mathbf{A} \leftarrow \left(\mathbf{X} + \frac{1}{\mu}\mathbf{Z}\right)\mathbf{B}\left(\mathbf{B}^T\mathbf{B} + \frac{\gamma}{\mu}\mathbf{I}\right)^\dagger, \quad (17)$$

$$\mathbf{B} \leftarrow \left(\mathbf{X} + \frac{1}{\mu}\mathbf{Z}\right)\mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \frac{\gamma}{\mu}\mathbf{I}\right)^\dagger. \quad (18)$$

The update of $\mathbf{X}$ is:

$$\mathbf{X} \leftarrow \mathcal{P}_\mathcal{A}\left(\frac{\mu}{\mu+2}\mathbf{A}\mathbf{B}^T - \frac{1}{\mu+2}(\mathbf{Z} - \mathbf{W}) + \frac{\rho}{\mu+2}\mathbf{Y}\right). \quad (19)$$

The solution for $\mathbf{X}$ turns out to be a projection to $\mathcal{A}$ and denotes by $\mathcal{P}_\mathcal{A}(\cdot)$.

[3]With the consideration that solving semidefinite programming is generally unscalable, we follow the previous works [60], [61] as ignoring the positive semidefinite constraint on $\mathbf{X}$ and give the low-rank constraint a large weight. Since the results are not degraded noticeably when removing the constraint on the sum of each row and column of $\mathbf{X}$, we remove it and preserve the requirement that elements in $\mathbf{X}$ should be lied in [0, 1], i.e., $0 \leq \mathbf{X} \leq 1$.

Algorithm 1 Co-Detection Algorithm

Input: Candidate similarity matrix $\mathbf{S}^v$, indication matrices $\mathbf{M}_i$'s, feature point locations $\mathbf{P}_i$'s, fundamental matrices $\mathbf{F}_{ij}$'s.
Output: Globally candidate correspondences $\mathbf{X}$.
Procedure:

```
Randomly initialize A and B, Z = 0, compute Y_ij's with Eq. 20;
Initialize X with Eq. 15 and set ρ = 0;
while not converged do
    while not converged do
        update A with Eq. 17;
        update B with Eq. 18;
        update X with Eq. 19;
        update Z with Z ← Z^k + μ(X − AB^T);
    update Y_ij's with the Hungarian algorithm in Eq. 20;
Quantize X with a threshold.
```

To update each $\mathbf{Y}_{ij}$, we utilize the Hungarian algorithm, with the constructed cost matrix as:

$$\mathbf{H}_{ij} = G(\mathbf{P}_i, \mathbf{P}_j) - \rho \mathbf{X}_{ij}, \tag{20}$$

where $G(\mathbf{P}_i, \mathbf{P}_j)$ is the geometrical response of the image pair (I_i, I_j). Elements in $\mathbf{G}$ are values denoting point correspondence with fundamental matrix between each pair of candidates.

As the optimization is non-convex and includes both continuous and discrete variables, the quality of the initialization of $\mathbf{X}$ is necessary. We first solve the Eq. 15 with $\rho = 0$ (without consideration of geometrical consistency) to obtain a confidential $\mathbf{X}$. The initialization of fundamental matrices $\mathbf{F}_{ij}$'s, is completed by RANSAC method with matching relations of feature points. With the estimated $\mathbf{F}_{ij}$'s, the assignments of $\mathbf{Y}_{ij}$'s are achievable.

For better convergence, based on the initialized $\mathbf{X}$ and $\mathbf{Y}$, we solve the Eq. 15 until reaching the local minimum and then update $\mathbf{Y}$. We iteratively conduct the updates to find the optimal $\mathbf{X}$. The overall algorithm is presented in the Algorithm 1.

IV. Experiments

In this section, we first present the new text co-detection dataset. Then we conduct experiments on the collected dataset to validate the effectiveness of the proposed text co-detection method. In addition, we investigate the performance and superiority of the co-detection framework comparing to several recent competing text detection methods in a single image.

A. Text Co-Detection Dataset

To evaluate the performance of text co-detection, we collect a new text dataset consisting of image groups each of which describes the same scene. Each image group holds several images photographed from the same scene. The dataset is collected from two sources. The first one includes frames extracted from videos. We choose 34 of 49 videos, which present diverse views of the scene, from the Text in Videos challenge of ICDAR 2019 Robust Reading Competition.[4] These videos are captured with a moving camera with changing camera locations, including angles, vertical moving, depth, etc. Texts in videos always suffer from natural noise, blurring, perspective distortion and substantial changes in illumination and occlusion. To select useful frames, we first pick frames with the interval of 20 frames, which may be enlarged to 50 frames when the content difference is visually small. Then the picked frames of each video are divided into several groups of frames, each of which describes the same scene from different views. Subsequently, for each group, we manually delete frames that are visually similar to the previous frames. Finally, the preserved groups constitute the part of the dataset. The second source for collecting image groups is images from the widely used text detection dataset ICDAR2015. We select images captured from the same scene as an image group. Texts in images from the ICDAR2015 dataset hold large-scale variances and are polluted by varying perspective, illumination and occlusion. Totally, the new dataset contains 240 groups with 727 images. We show two groups deriving from different sources in Fig. 5. In Fig. 5 (a), it is obvious that text "Aparcament", "reservat" and "Rectorat" become more and more clear and are easy to capture from the top left to the down right in the zigzag direction. Therefore, incorporating the text detection results of the down right image benefits the detection results of the top left image. Besides, since the left top image is captured with deep depth, it may provide the global view to locate texts in the scene. In Fig. 5 (b), horizontal texts in the down image may be easier to obtain compared with the tilted and depth-affected texts in the top image, further provide guidance to assist text detection in the top image.

(a) (b)

Fig. 5. Examples of the collected text co-detection dataset. The group (a) comes from the video source and the group (b) from the image source.

Since having a consistent and reliable groundtruth is imperative to carrying out an evaluation, we consider scene characteristics such as spatial resolution of texts and corresponding persistence of texts to decide the groundtruth annotation way. Firstly, given an image set, areas that covering readable texts, words, text lines, partial letters and unclear text lines are marked with four points in the clockwise direction, i.e., locations at top left, top right, down right and down left. Then we

[4]The challenge is available at https://rrc.cvc.uab.es/?ch=3

select boxes that cover the same scene text and present in two or more images as the corresponding text annotations.

B. Evaluation and Comparison

To evaluate the text co-detection performance, we approach with a spatio-temporal concept for the correspondence measure in comparison with the text detection measure that taking spatial aspect into account. Before we conduct the performance evaluation, we require to transform the correspondence representation matrix $\mathbf{X}$ to a set of texts with correspondences and assign each valid detection with a groundtruth index. For each candidate in $\mathbf{X}$, if it has a valid intra-image correspondence, it is treated as a co-text candidate. For example, for the candidate c_i^m in image I_i, if $x_{i*}^{m*} >= 0.5$ ($*$ indicates any text candidate in any image except for I_i), the candidates c_i^m and c_*^*s compose a co-text candidate. For each co-text candidate, a one-to-one correspondence between the detected co-text and the groundtruth is determined by filtering the Intersection over Union (IoU) score with 0.5 and selecting the maximal IoU score over all combinations of the detection and the groundtruth. The remained co-text candidates are detected co-texts. The following are the notations used in the performance evaluation,

- G_s is the s^{th} groundtruth text and G_s^i denotes the s^{th} groundtruth text in image I_i.
- D_s is the s^{th} detected co-text and D_s^i denotes the s^{th} detected co-text in image I_i.
- N_G^i and N_D^i are the number of groundtruth texts and the number of detected co-texts in image I_i, respectively.
- N_G and N_D are the number of unique groundtruth texts and the number of detected co-texts in the given image set, respectively.
- N is the number of images in a given image set and N_{corr} is the number of detected co-texts.

We first employ the detection accuracy (Image Detection Accuracy (IDA)) and recall (Image Detection Recall (IDR)) measures to estimate the image-level co-detection performance of detected co-texts. These two measures are defined as,

$$IDA = \sum_{i=1}^{N} \frac{\sum_{s=1}^{N_{corr}} |D_s^i|}{N_D^i}. \tag{21}$$

$$IDR = \sum_{i=1}^{N} \frac{\sum_{s=1}^{N_{corr}} |D_s^i|}{N_G^i}. \tag{22}$$

We then propose a Multiple Image Detection Ratio Accuracy (MIDRA), which is an image-level measure and accounts for number of texts detected, missed detection, and false alarms of groundtruth and discovered texts. It sequentially computes detection ratio for each image and then normalizes all IDRA scores to obtain the performance of multiple images. The Image Detection Ratio Accuracy (IDRA) for image I_i is defined as,

$$IDRA(i) = \frac{2 \times \sum_{s=1}^{N_{corr}} \frac{|G_s^i \cap D_s^i|}{|G_s^i \cup D_s^i|}}{N_G^i + N_D^i}, \tag{23}$$

where $\frac{|G_s^i \cap D_s^i|}{|G_s^i \cup D_s^i|}$ is the IoU score between the s^{th} detected text bounding box for image I_i and the groundtruth text bounding box.

To obtain the MIDRA performance, the IDRA scores for each image are summed together and normalized by the number of images that either holds a groundtruth or a detected text. The formula is expressed as,

$$MIDRA = \frac{\sum_{i=1}^{N} IDRA(i)}{\sum_{i=1}^{N} \exists(N_G^i \ OR \ N_D^i)}. \tag{24}$$

In spite of MIDRA that considers image-level co-detection performance, we additionally employ another evaluation criteria, the Average Correspondence Accuracy (ACA), to estimate the correspondence accuracy performance for each co-text. Inspired by [67], we first calculate the Correspondence Detection Accuracy (CDA), which shows the performance of one co-text on all images. The CDA is defined as,

$$CDA = \sum_{s=1}^{N_{corr}} \frac{\sum_{i=1}^{N} [\frac{|G_s^i \cap D_s^i|}{|G_s^i \cup D_s^i|}]}{N_{(G_s \cup D_s \neq \emptyset)}}. \tag{25}$$

Then the ACA is obtained by normalizing the CDA with the average number of texts in the image set. It is defined as,

$$ACA = \frac{2 \times CDA}{N_G + N_D}. \tag{26}$$

Among the above proposed four evaluation metrics, IDA and IDR measure the accuracy and recall, respectively, of detected co-texts. They consider the number of texts being correctly detected and evaluate the performance by calculating the ratio of the number of the correctly detected texts to that of detected texts or groundtruth texts. These two metrics are similar with the Accuracy and Recall metrics in the single image text detection evaluation, and ignores the evaluation of correspondence relationships. Therefore, correspondence errors are not presented in these two metrics.

Compared to IDA and IDR, MIDRA and ACA provide more comprehensive analysis about text co-detection. MIDRA fully considers number of detected texts, missed texts, and false alarms of groundtruth and discovered texts. It calculates the multiple image performance by normalizing the image-level ratio accuracy with the total number of images that either has a groundtruth or a detected object. This normalization considers both missed detection and false alarms. ACA calculates the accuracy based on each corresponding text set. It covers incorrect text propagation along co-texts.

We extract candidate deep feature representation from the last 4^{th} convolutional feature responses of the SSTD model via Region of Interest (RoI) pooling strategy. The intra-image candidate similarities are ignored by assigning with score 0 and the self-similarities of candidates are assigned with 1, i.e., diagonal elements in similarity matrices are equal to 1. The weight of geometrical consistency β is assigned with 0.01, the weight of nuclear norm γ with 50, and ρ with 1.

We construct the comparison with frameworks that have the same goal to explore relative and consistent components among multiple images, which aim to explore common objects and their relationships. We also show the result of the proposed

method compared with the combination of the-state-of-the-art text detection methods and two relationship discovery frameworks, which have the ability to perform co-texts. The comparison methods are manifested in the following.

1) CTPN [34] detects text lines based on sequential text proposals in convolutional feature maps. We use the released model and preserve the parameter configuration presented in the paper.[5]

2) TextBoxes [31] presents an end-to-end trainable scene text detector without post-process except for a standard non-maximum suppression.[6]

3) SSTD [30] employs an attention mechanism which roughly identifies text regions via an automatically learned attentional map.[7]

4) TextPros [68] generates a hierarchy of word hypotheses that rely on a similarity based region grouping.[8]

5) EAST [32] directly predicts texts with arbitrary orientations and quadrilateral shapes based on a fully convolutional network.[9]

6) SegLink [29] decomposes text into two locally detectable elements, namely segments and links, and produces texts by combining segments connected by links. We use the model trained using the image size of 384×384, and set thresholds for the confidence of segments and the confidence of linking as 0.9 and 0.7, respectively, all of which are same with those described in the paper.[10]

7) RRPN [69] proposes to utilize inclined proposals with text orientation information to conduct the text detection task.[11]

8) PixelLink [70] is constructed based on instance segmentation. It first segments out text instances via linking pixel-level neighborhoods within the same instance and then generates text bounding boxes from segmented text instances. It considers pixel-level links and formulates connected components as detected texts.[12]

9) Textspotter [36] utilizes a four-components framework to classify and segment text proposals and deal with arbitrary-oriented scene text detection. They consider both global word map and character maps to provide accurate localization.[13]

10) PW(PairWise similarity) is completed only based on pairwise similarities and the greedy algorithm to find candidate pairs holding the largest similarities. Similarities between candidates from the same image are assigned with 0. For each image pair, we employ the pair-wise similarity maximization strategy to find the correspondences. The correspondences for multiple images are determined via maximizing a sequence of pair-wise correspondences. Then the detected corresponding co-texts are assigned with the groundtruth ID that has largest IoU score.

11) MSG [40] presents a co-segmentation framework for multiple foreground video co-segmentation in a set of videos, which has similar goal with our proposed co-detection framework for multi-view text detection. It considers coherences of the foreground within the video and among the different videos.

12) Clique [41] proposes a multi-image co-segmentation framework. It constructs the graph structure to represent the candidate relationships among foreground candidates and seeks cliques linking candidates coming from multiple images as desired foregrounds.

13) Ours$_{w/o}$ is the proposed text co-detection framework without considering the geometrical guidance. It is implemented based on the Eq. 11 that removes the second term $\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\|\hat{\mathbf{P}}_i\mathbf{F}_{ij}\hat{\mathbf{P}}_j\|_F^2$ and also solved with the same strategy ADMM.

The experimental results are shown in Table II. We first show the results of the combination of the text detection methods and two correspondence exploration methods. For the first combination, we employ the pair-wise similarity maximization strategy (PW) after obtaining text detection results to conduct correspondence discovery. For the second combination, we incorporate the text detection results and our proposed framework without the geometrical guidance (Ours$_{w/o}$). The results are presented in the first and second blocks. We observe that incorporating the text detection methods with our proposed framework Ours$_{w/o}$ has superior performance than that with the PW strategy. Taking the results of SSTD as example, which constitutes part of our candidate set. We compare the performance of SSTD+PW, SSTD+Ours$_{w/o}$ and Ours$_{w/o}$. Co-texts generated by SSTD+PW only obtain 11.75% and

TABLE II

PERFORMANCES OF DIFFERENT METHODS ON THE TEXT CO-DETECTION DATASET

Methods \ Metrics	IDA	IDR	MIDRA	ACA
CTPN [34] + PW	6.91%	8.68%	5.78%	8.44%
TextBoxes [31] + PW	8.88%	11.77%	8.60%	11.19%
SSTD [30] + PW	11.75%	15.07%	10.87%	14.98%
TextPros [68] + PW	7.57%	8.68%	7.46%	8.89%
SegLink [29] + PW	11.92%	14.43%	10.12%	14.21%
EAST [32] + PW	11.74%	14.17%	9.75%	14.17%
RRPN [69] +PW	12.36%	15.17%	10.71%	15.52%
PixelLink [70] +PW	13.99%	16.90%	12.07%	17.22%
TextSpotter [36] +PW	17.31%	19.47%	14.94%	20.82%
CTPN [34] + Ours$_{w/o}$	11.35%	13.93%	11.44%	13.25%
TextBoxes [31] + Ours$_{w/o}$	9.01%	11.68%	10.02%	11.00%
SSTD [30] + Ours$_{w/o}$	20.03%	22.28%	20.68%	21.14%
TextPros [68] + Ours$_{w/o}$	12.90%	13.60%	17.29%	13.16%
SegLink [29] + Ours$_{w/o}$	18.98%	21.27%	18.61%	21.22%
EAST [32] + Ours$_{w/o}$	20.49%	23.66%	20.84%	23.44%
PixelLink [70] +Ours$_{w/o}$	21.47%	24.97%	23.52%	24.82%
RRPN [69] +Ours$_{w/o}$	18.28%	21.57%	20.07%	21.64%
TextSpotter [36] +Ours$_{w/o}$	22.43%	23.61%	22.97%	26.04%
PW	6.93%	8.28%	6.62%	8.72%
MSG [40]	5.62%	6.05%	7.63%	5.97%
Clique [41]	8.53%	9.74%	10.12%	9.05%
Ours$_{w/o}$	22.07%	23.73%	22.94%	24.11%
Ours	**25.41**%	**27.66**%	**26.93**%	**28.32**%

[5]The code and model are available at https://github.com/tianzhi0549/CTPN.git

[6]The released code is available at https://github.com/MhLiao/TextBoxes.git

[7]The code is available at https://github.com/BestSonny/SSTD

[8]The model is available at https://dl.dropboxusercontent.com/u/45812668/dictnet_vgg/dictnet_vgg.caffemodel

[9]The code is available at https://github.com/argman/EAST

[10]The details for code and model are available at https://github.com/dengdan/seglink/introduction

[11]The model is available at https://github.com/mjq11302010044/RRPN

[12]The model is available at https://github.com/ZJULearning/pixel_link

[13]The model is available at https://github.com/MhLiao/MaskTextSpotter

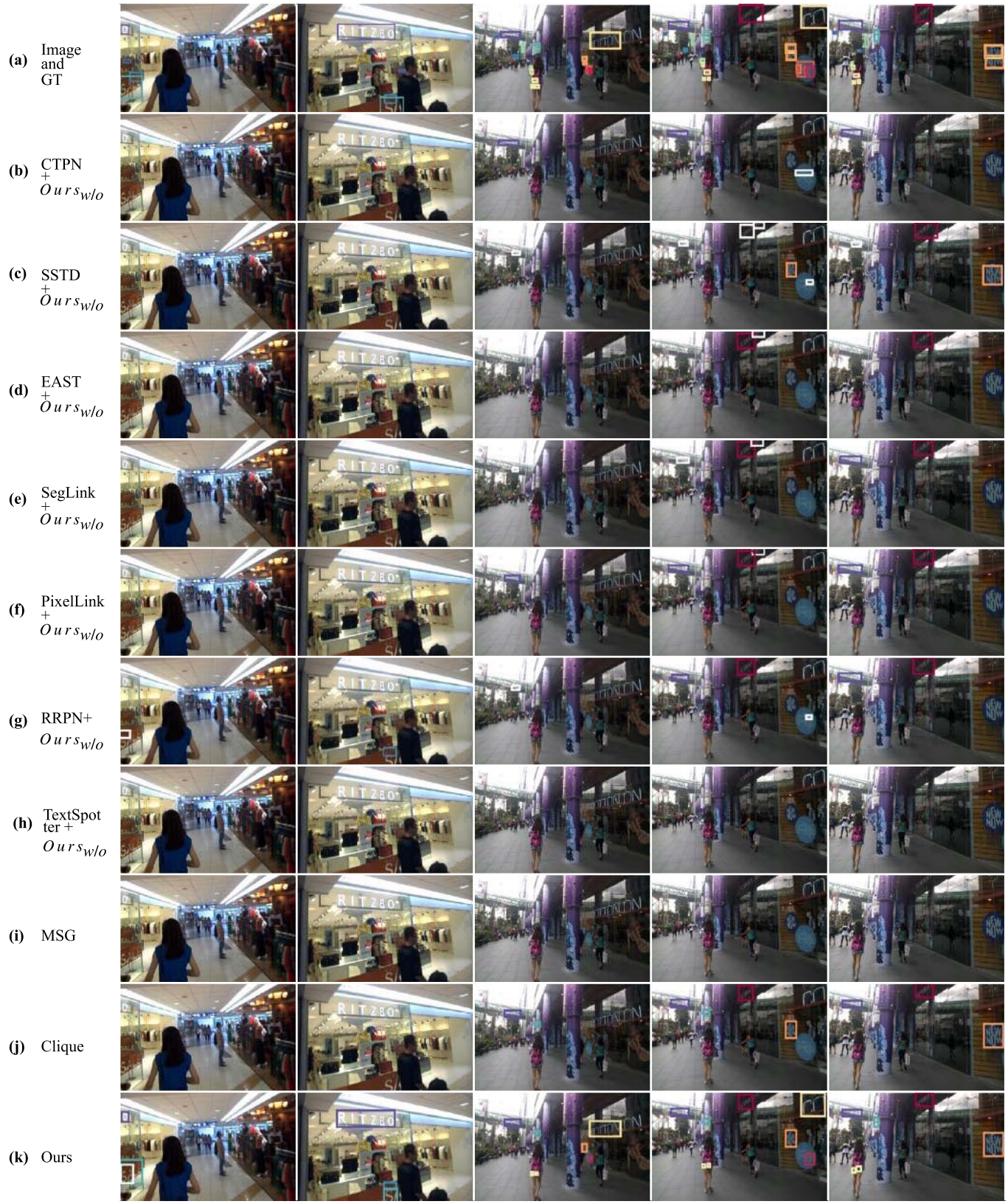


Fig. 6. Qualitative results for text co-detection. Each row indicates the co-occurring text groundtruth (a), CTPN+$Ours_{w/o}$ (b), SSTD+$Ours_{w/o}$ (c), EAST+$Ours_{w/o}$ (d), SegLink+$Ours_{w/o}$ (e), PixelLink+$Ours_{w/o}$ (f), RRPN+$Ours_{w/o}$ (g), TextSpotter+Ours$_{w/o}$ (h), MSG (i), Clique (j) and Our results (k), respectively. Bounding boxes with the same color indicate one co-text. The white rectangle is the output detection whose IoU score is less than 0.5. Our proposed co-detection method jointly capture corresponding text detection, and discover relational correspondences.

15.07% on IDA and IDR, respectively. When accomplishing the one-to-one constraint and cycle consistency, the IDA and IDR have the increase of 70.5% and 47.8%, respectively. But it is obvious that although we have the-state-of-the-art text

detection methods like EAST and good correspondence discovery strategy, the performance is still inferior compared with the proposed text co-detection frameworks Ours$_{w/o}$ and Ours. When conducting detection and matching together, the IDA and IDR receive the additional improvements of 17.4% and 9.6%, respectively. The higher improvements are also presented on the performance of MIDRA and ACA metrics. These comparisons show that our proposed text co-detection framework, which jointly conducts text detection and correspondence discovery, has high effectiveness on performing co-text discovery.

Further, we also show the results of co-frameworks, which are conducted on the multi-view dataset, in the third block in Table II. From results in the table, we find that co-segmentation frameworks seem to perform worse. It possibly due to the reason that the co-segmentation framework aims to discover salient and well-defined objects as foregrounds, which may perform the weak ability on cluttered and not-well-defined text regions.

We present some qualitative results in Figure 6 to show the detected texts and the explored text correspondences. Texts framed with the same color indicate the co-texts. Bounding boxes with white color are results whose IoU are less than 0.5. It is straightforward that our proposed text co-detection framework not only has good text detection results but also is well-behaved on finding correspondences of detected texts. Compared with combinations of single-view text detection methods and correspondence discovery strategies, the text co-detection results of our proposed multi-view co-detection framework are obviously superior. Compared with co-frameworks MSG [40] and Clique [41], our proposed multi-view co-detection framework obtains better text detection and correspondence exploration results. Specifically, the MSG [40] requires the assignment of the number of co-texts, which is impractical since the number of co-texts are uncertain. In practice, we set the number of co-texts with 15 and preserve correspondences that are unique and presented in the first time. The co-framework Clique [41] sequentially seeks exact one-to-one relationships between candidates among images, and ignores the relationships among images.

In summary, compared with other methods, the proposed method not only consider the global consistency but also the local consistency. For the global consistency, the cycle consistency prevents mismatches among multiple images and ensures candidate selections that are incorrect in paired images. For the local consistency, the employed epipolar-geometry information explores relationships of text-specific local feature points and further preserves correspondences of text candidates.

C. Analysis and Ablation Study

In this section, we perform the analysis about the effectiveness of features and terms employed in the text co-detection framework.

1) Analysis of Candidate Selection and SWT Region Generation: We first analyze the thresholds used in candidate generation step. In this paper, text candidates are generated based on deep-learning representations and text-specific features. We use SSTD [30] network to conduct candidate search, where each candidate has a probability score to measure whether the candidate being a text region. Instead of using a higher score to filter regions with low probability, we relax the high probability requirement and reduce the threshold to include all possible text regions. In Table III, we present the Recall evaluation, which is designed to measure the number of ground-truth texts being detected, and the number of generated candidates against different confidence thresholds on the ICDAR 2015 testing dataset. We choose 0.3 as the desired confidence threshold since Recall under 0.3 is sufficiently high and the number of candidates is not very large.

TABLE III
RECALL OF TEXT CANDIDATES WITH DIFFERENT CONFIDENCE THRESHOLDS ON ICDAR 2015 TESTING DATASET

Thresh	0.6	0.5	0.4	0.3	0.2
Recall	42.16 %	45.23%	48.66%	52.53%	57.25%
number of Candidates	6	7	9	12	29

Considering the text-specific features, we use SWT interesting points [18] to describe text-aware locations and group neighboring SWT points as a text region. Since SWT-based text regions are derived from point locations, there may exist little regions having few text information and visually unnoticeable. We conduct a reduction progress to remove text regions whose width or height is less than 10 pixels. This threshold is designed based on two considerations. The first one is based on the statistical analysis about size of text regions of ICDAR 2015 text detection dataset and Texts in Videos dataset. We calculate the widths and heights of all ground-truth text regions and find that the minimal width and height are higher than 5 pixels. The second one is based on the observation that text regions whose width or height is smaller than 10 pixels are always visually unclear. Therefore, we discard SWT regions by filtering their widths and heights with a threshold.

2) Effectiveness of Different Feature Representations: Given the extracted text candidate locations, we extract visual feature representations from the SSTD deep model with 5 convolutional layers. We verify the representation ability of different convolutional layers, excluding influence from the geometrical guidance, and show the performance in Table IV. The results show that deeper layers have a stronger ability to represent texts. That is, features from the 4^{th} and 5^{th} layers produce superior performance than that from the 3^{th} layer. This is mostly due to the fact that a deeper representation has stronger representation ability to describe images, compared with shallower representation, since deeper representation captures not only visual information but also the semantic representation, e.g., text and non-text. However, the representations from the 5^{th} layer do not hold better results than that from the 4^{th} layer. It is probably induced by two reasons. Firstly, since text regions are smaller, after several down-sampling operations, some neighboring text regions may have the same responses on the 5^{th} layer. Besides, features from the 5^{th} layer have weak visual description ability and may pay more attention to judging the probability of being a text region, hence they have weak discriminative ability between text candidates.

TABLE IV
PERFORMANCE WITH DIFFERENT DEEP REPRESENTATIONS

Metrics \ Feature Layer	IDA	IDR	MIDRA	ACA
$conv_3$	20.17%	23.05%	20.82%	21.46%
$conv_4$	**25.41%**	**27.66%**	**26.93%**	**28.32%**
$conv_5$	21.03%	23.74%	21.29%	22.63%

TABLE V
ABLATION STUDY

Method \ Metric	IDA	IDR	MIDRA	ACA
PW_w	9.75%	10.37%	10.16%	10.29%
$Ours_{w/o}$	22.07%	23.73%	22.94%	24.11%
Ours	**25.41%**	**27.66%**	**26.93%**	**28.32%**

3) Analysis on Complexity and Speed: We first analyze the time complexity about the proposed algorithm described in Algorithm 1. The complexity of the framework is $O(M(Q^3r^4+Q^2+P^2))$, where Q is the number of candidates of all images, P is the number of interest points of all images. r (smaller than Q) is a dimension of A and B and M is the iteration number. Besides, we also compute the time consumption and the average processing time for each image group (including more than two images) without the consideration of the candidate generation step. The average time consumption of one image group is 2.7s under a MacBook Pro with 2.7 GHz Intel i5 Core and 16GB DDR3 memory.

4) Analysis on the Effectiveness of Correspondences to Text Detection: In the single-view detection, the detector conducts the prediction based on the probability score of being a text. A text region having high score is treated as positive and the text region having low score is negative. The threshold for deciding to be a text or not plays important role in text detection. However, since texts are extremely sensitive to environmental conditions like illumination, capture angle like perspective, camera blur and etc., the same text under different environments may have different probability scores lower or higher than the threshold. Thus the texts with lower scores are prevented from being detected.

Different with the single-view text detection, the proposed text co-detection not only considers probability scores of texts, but also corresponding relationships between texts in different view. With the assistance of corresponding relationship, texts with lower scores still have opportunity to be detected. Therefore, the relationship assists the improvement of detection accuracy.

In practice, it is inevitable that there exist false relationships in multiview group. However, the quantitive comparisons show that the corresponding relationships still have strong positive effectiveness on improving the detection performance. Taking the results of SSTD and ours shown in Table II as example, correspondences explored with the one-to-one constraint and the cycle consistency get 70.5% and 47.8% improvements on IDA and IDR, separately, compared to pairwise similarity relationships (the first row). Furthermore, with the help of the geometrical relationships, the new correspondences bring the further improvements. All the comparison results demonstrate the benefits of corresponding relationships.

5) Ablation Study: To take the consistency among multiple images into consideration, we enforce the cycle consistency for discovered co-occurring text regions. To evaluate the effectiveness of the cycle consistency, the comparison between a baseline PW_w method and our proposed text co-detection framework is shown in the Table V. The PW_w method is implemented based on the union of the visual similarities and geometrical guidance of text candidates. The results present a clear illustration about the influence of the cycle consistency.

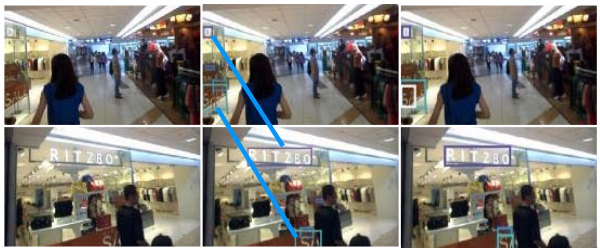


Fig. 7. (a) The text co-detection by $Ours_{w/o}$. (b) The geometrical relationship (black line) obtained by epipolar geometry. (c) The co-texts detected by the proposed co-detection algorithm.

In spite of the cycle consistency, we also put the intrinsic geometrical correspondence of multi-view images, which described as the epipolar guidance, as the additional information for text co-detection. The geometrical guidance directly requires the matrix $\mathbf{X}$, which indicates text candidate selection, having correspondences with the selection matrix $\mathbf{Y}$ calculated based on epipolar geometrical relation. To perform the effects of the added geometrical guidance, we conduct the comparison baseline $Ours_{w/o}$. The comparisons are shown in Table V. The superior performance shows that the cycle consistency requirement benefits the co-detection results.

The epipolar geometrical guidance also has the positive influence on text co-detection. By comparing with the results of Ours and $Ours_{w/o}$, it is obvious that geometrical relationship benefits the performance of text co-detection. Considering that merely exploring correspondences based on visual feature representations may be misguided since texts are sensitive to environmental conditions, incorporating the relationships derived from visual features with that from geometrical guidance largely improves the performance.

To further analysis the employed geometrical guidance, we show a qualitative example showing the comparison between detected co-texts by $Ours_{w/o}$ and Ours, with the presence of the geometrical guidance of the image pair. The results are shown in Fig. 7. Since co-texts in these two images have large visual differences, merely using similarity-based relationships could provide limited efforts to detect co-texts, such as results in the first column in Fig. 7. After integrating geometrical guidance (shown as the black lines cross top and down images in the second column in Fig. 7), we obtain satisfied co-text detection results, as shown in the third column in Fig. 7.

We owe the success of adding geometrical guidance to the following three aspects. Firstly, the geometrical guidance is constructed based on the relationships of text-specific

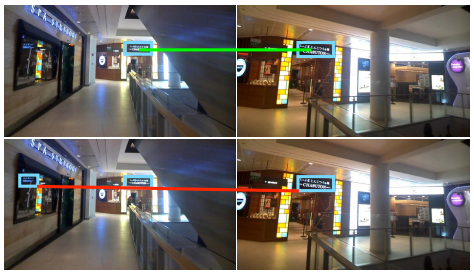


Fig. 8. Incorrect candidate correspondence caused by reflection.

SWT points. The number of SWT points in each image is more than 2,000 and thus provides abundant sources to estimate the geometrical relationships. Secondly, given the estimated SWT point pairs, we use the voting strategy to determine the final geometrical guidance between candidates. The voting strategy, which counts the number of SWT point pairs contributing to the candidates, integrates the point-level relationships to candidate-level and thus may avoid the distraction induced by the minority. Thirdly, as described in Algorithm 1, the geometrical guidance directly attends the update of the desired $\mathbf{X}$, instead of being a filter after $\mathbf{X}$ is obtained. This incorporation makes the algorithm sufficiently and simultaneously consider similarity relationships and geometrical guidance, thus brings the improvements.

D. Failure Case Analysis

Although the proposed text co-detection handles texts that are difficult to be localized due to large variations of perspective, illumination, occlusion and etc., but there are some special situations that cannot be tackled. As shown in Fig. 8, green link in the top row shows the correct correspondence between two candidates and red link in the bottom row is the wrong correspondence. The reason of this wrong correspondence is that the candidate links to the reflection region of its corresponding candidate. The reflection region has higher visual feature similarity than the original candidate, with the candidate in the right.

V. Conclusion

In this paper, we have focused on a new task, text co-detection, which aimed to detect the corresponding texts and identify their correspondences from the multi-view scene with large perspective and environmental condition variations. To deal with this task, we have proposed a novel text co-detection framework, which integrated the visual feature similarities and geometrical guidance. The cycle consistency constraint has also been incorporated to ensure the relationships of texts among multiple images. Moreover, a text co-detection dataset has been collected for evaluation.

References

[1] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki, "GIFT: A real-time and scalable 3D shape search engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5023–5032.

[2] E. Johns, S. Leutenegger, and A. J. Davison, "Pairwise Decomposition of Image Sequences for Active Multi-view Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3813–3822.

[3] T. Yu, J. Meng, and J. Yuan, "Multi-view harmonized bilinear network for 3D object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 186–194.

[4] A. Kanezaki, Y. Matsushita, and Y. Nishida, "RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5010–5019.

[5] T. Tlusty, T. Michaeli, T. Dekel, and L. Zelnik-Manor, "Modifying non-local variations across multiple views," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6276–6285.

[6] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 406–413.

[7] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2821–2830.

[8] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1945–1954.

[9] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6489–6498.

[10] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, Jun. 2017, pp. 5028–5037.

[11] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3980–3989.

[12] H. Rhodin *et al.*, "Learning monocular 3D human pose estimation from multi-view images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8437–8446.

[13] T. Isokane, F. Okura, A. Ide, Y. Matsushita, and Y. Yagi, "Probabilistic plant modeling via multi-view image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2906–2915.

[14] C. Jung, Q. Liu, and J. Kim, "A stroke filter and its application to text localization," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 114–122, Jan. 2009.

[15] Y. Gao, Y. Chen, J. Wang, M. Tang, and H. Lu, "Reading scene text with fully convolutional sequence modeling," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 3501–3508.

[16] M. Tanaka and H. Goto, "Autonomous text capturing robot using improved DCT feature and text tracking," in *Proc. 5th Int. Conf. Doc. Anal. Recognit.*, Paraná, Brazil, Sep. 2007, pp. 1178–1182.

[17] L. Gomez and D. Karatzas, "MSER-based real-time text detection and tracking," in *Proc. Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 3110–3115.

[18] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2963–2970.

[19] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018.

[20] S. Tian, X.-C. Yin, Y. Su, and H.-W. Hao, "A unified framework for tracking based text detection and recognition from Web videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 542–554, Mar. 2018.

[21] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comput. Vis.*, Queenstown, New Zealand, Nov. 2010, pp. 770–783.

[22] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1241–1248.

[23] K. Wang and S. J. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 591–604.

[24] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 97–104.

[25] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, Sep. 2015.

[26] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.
[27] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 800–813, Mar. 2011.
[28] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 2609–2612.
[29] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3482–3490.
[30] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 3066–3074.
[31] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proc. 31th AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4161–4167.
[32] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2642–2651.
[33] S. Zhu and R. Zanibbi, "A text detection system for natural scenes with convolutional feature learning and cascaded classification," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 625–632.
[34] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 56–72.
[35] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.
[36] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 71–88.
[37] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5909–5918.
[38] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A Flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 19–35.
[39] R. Quan, J. Han, D. Zhang, and F. Nie, "Object co-segmentation via graph optimized-flexible manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 687–695.
[40] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.
[41] C. Wang, H. Zhang, L. Yang, X. Cao, and H. Xiong, "Multiple semantic matching on augmented N-partite graph for object co-segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5825–5839, Dec. 2017.
[42] H. Liu, Z. Tao, and Y. Fu, "Partition level constrained clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2469–2483, Oct. 2018.
[43] Y. Ren, L. Jiao, S. Yang, and S. Wang, "Mutual learning between saliency and similarity: Image cosegmentation via tree structured sparsity and tree graph matching," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4690–4704, Sep. 2018.
[44] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[45] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 38:1–38:31, 2018.
[46] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.
[47] Z. Hayder, X. He, and M. Salzmann, "Structural kernel learning for large scale multiclass object co-detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 2632–2640.
[48] V. G. Kim, W. Li, N. J. Mitra, S. DiVerdi, and T. A. Funkhouser, "Exploring collections of 3d models using fuzzy correspondences," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 54:1–54:11, 2012.
[49] Q. Huang, F. Wang, and L. Guibas, "Functional map networks for analyzing and exploring large shape collections," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 36:1–36:11, 2014.
[50] S. Y. Bao, Y. Xiang, and S. Savarese, "Object co-detection," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 86–101.
[51] J. Shi, R. Liao, and J. Jia, "CoDeL: A human co-detection and labeling framework," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2096–2103.
[52] Z. Hayder, M. Salzmann, and X. He, "Object co-detection via efficient inference in a fully-connected CRF," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 330–345.
[53] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1228–1242, Jun. 2016.
[54] T. Zhou, Y. J. Lee, S. X. Yu, and A. A. Efros, "FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1191–1200.
[55] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[56] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[57] D. Pachauri, R. Kondor, and V. Singh, "Solving the multi-way matching problem by permutation synchronization," in *Proc. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 1860–1868.
[58] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M. Chu, "Graduated consistency-regularized optimization for multi-graph matching," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 407–422.
[59] Y. Chen, L. J. Guibas, and Q. Huang, "Near-optimal joint object matching via convex relaxation," in *Proc. Int. Conf. Mach. Learn.* Beijing, China, Jun. 2014, pp. 100–108.
[60] Q.-X. Huang and L. Guibas, "Consistent shape maps via semidefinite programming," *Comput. Graph. Forum*, vol. 32, no. 5, pp. 177–186, Aug. 2013.
[61] X. Zhou, M. Zhu, and K. Daniilidis, "Multi-image matching via fast alternating minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 4032–4040.
[62] N. Parikh and S. P. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.
[63] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[64] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *Soc. Ind. Appl. Math. Rev. J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.
[65] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Dec. 2013, pp. 2488–2495.
[66] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *Soc. Ind. Appl. Math. Rev.*, vol. 52, no. 3, pp. 471–501, Jan. 2010.
[67] V. Manohar *et al.*, "Performance evaluation of text detection and tracking in video," in *Proc. 7th Int. Workshop Document Anal. Syst.*, Nelson, New Zealand, Feb. 2006, pp. 576–587.
[68] L. Gómez and D. Karatzas, "TextProposals: A text-specific selective search algorithm for word spotting in the wild," *Pattern Recognit.*, vol. 70, pp. 60–74, Oct. 2017.
[69] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[70] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Apr. 2018, pp. 6773–6780.

Chuan Wang is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her current research interests include co-segmentation, co-detection, matching, and people counting.

Huazhu Fu (Senior Member, IEEE) received the Ph.D. degree in computer science from Tianjin University, China, in 2013. He was a Research Fellow with Nanyang Technological University, Singapore, for two years. From 2015 to 2018, he was a Research Scientist with Institute for Infocomm Research, Agency for Science, Technology, and Research, Singapore. He is currently a Senior Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, image processing, and medical image analysis. He is an Associate Editor of the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, the IEEE ACCESS, and *BMC Medical Imaging*.

Liang Yang received the B.E. and M.E. degrees in computational mathematics from Nankai University, Tianjin, China, and the Ph.D. degree in computer science from the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Technology, Tianjin. His current research interests include community detection, graph neural networks, low-rank modeling, and data mining.

Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, USA. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, since 2012. He is also with Peng Cheng Laboratory, Cyberspace Security Research Center, China, and the School of Cyber Security, University of Chinese Academy of Sciences, China. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award from the International Conference on Pattern Recognition. He is on the Editorial Boards of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.