Social Context-aware Person Search in Videos via Multi-modal Cues

DAN LI, TONG XU, PEILUN ZHOU, and WEIDONG HE, University of Science and Technology of China, China
YANBIN HAO, City University of Hong Kong, China
YI ZHENG, Huawei Technologies, China
ENHONG CHEN, University of Science and Technology of China, China

Person search has long been treated as a crucial and challenging task to support deeper insight in personalized summarization and personality discovery. Traditional methods, e.g., person re-identification and face recognition techniques, which profile video characters based on visual information, are often limited by relatively fixed poses or small variation of viewpoints and suffer from more realistic scenes with high motion complexity (e.g., movies). At the same time, long videos such as movies often have logical story lines and are composed of continuously developmental plots. In this situation, different persons usually meet on a specific occasion, in which informative social cues are performed. We notice that these social cues could semantically profile their personality and benefit person search task in two aspects. First, persons with certain relationships usually co-occur in short intervals; in case one of them is easier to be identified, the social relation cues extracted from their co-occurrences could further benefit the identification for the harder ones. Second, social relations could reveal the association between certain scenes and characters (e.g., classmate relationship may only exist among students), which could narrow down candidates into certain persons with a specific relationship. In this way, high-level social relation cues could improve the effectiveness of person search. Along this line, in this article, we propose a social context-aware framework, which fuses visual and social contexts to profile persons in more semantic perspectives and better deal with person search task in complex scenarios. Specifically, we first segment videos into several independent scene units and abstract out social contexts within these scene units. Then, we construct inner-personal links through a graph formulation operation for each scene unit, in which both visual cues and relation cues are considered. Finally, we perform a relation-aware label propagation to identify characters' occurrences, combining low-level semantic cues (i.e., visual cues) and high-level semantic cues (i.e., relation cues) to further enhance the accuracy. Experiments on real-world datasets validate that our solution outperforms several competitive baselines.

CCS Concepts: • **Information systems**; • **Computing methodologies**;

Additional Key Words and Phrases: Person search, graph modeling, user profile, label propagation, social relation, neural network

This work was partially supported by the grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (No. 62072423).

Authors' addresses: D. Li, T. Xu (corresponding author), P. Zhou, W. He and E. Chen, University of Science and Technology of China, Hefei, Anhui, China; emails: lidan528@mail.ustc.edu.cn, {tongxu, cheneh}@ustc.edu.cn; Y. Hao, City University of Hong Kong, China; email: haoyanbin@hotmail.com; Y. Zheng, Huawei Technologies, Hangzhou, Zhejiang, China; email: zhengyi29@huawei.com.

1046-8188/2021/11-ART52 $15.00
https://doi.org/10.1145/3480967

ACM Reference format:
Dan Li, Tong Xu, Peilun Zhou, Weidong He, Yanbin Hao, Yi Zheng, and Enhong Chen. 2021. Social Context-aware Person Search in Videos via Multi-modal Cues. *ACM Trans. Inf. Syst.* 40, 3, Article 52 (November 2021), 25 pages.
https://doi.org/10.1145/3480967

1 INTRODUCTION

Recent years have witnessed the development of video-based person retrieval techniques, which has been becoming increasingly needed in real-world applications. For example, in law enforcement, to identify a specific criminal in a series of surveillance videos, huge amounts of human resources and time will be spent without automatic and high-quality video retrieval tools. Meanwhile, the boom of online sharing media contents also raises various application scenarios for video retrieval. For instance, audiences are usually keen on their favorite stars in movies or TV series, producing huge business values for video platforms to generate character summarization or reference photos for popular stars, and recommend these contents to corresponding users. In this way, among some of the biggest video platforms, the function of "Only look at him" has been designed for those fans who enjoy the video summarization of thier idols. Moreover, person search in videos also serves as a fundamental task for various downstream analysis tasks, such as the media content recommendation and the character modeling as well as the establishment of multi-modal knowledge graph. Therefore, it is significant to explore and enhance the video-based person retrieval system for videos with complicated scenes.

A few approaches have been proposed to tackle the person retrieval problem, such as Person Re-identification [8, 11, 27] and the Person Recognition [7, 40]. However, these traditional methods that use visual features to profile a person are not good at dealing with long videos due to the following reasons: First, the traditional visual-based approaches are usually conducted in monitoring scenes, in which query and gallery images share relatively easy conditions, which suffer from the dramatic variations in clothing, pose, and even makeup in long videos. Second, these traditional methods neglect rich semantic information in complex videos. Specifically, complex videos like movies usually have a logical and descriptive story line, containing not only visual information but also semantic cues , which could help to reveal social relations between characters and assist to profile characters more accurately. With problems above, it remains challenging to tackle the person retrieval well in long videos with plot lines.

Along this line, to identify target characters in complex videos, we propose to utilize their social relations revealed by multi-modal contextual information. Intuitively, the easily recognized character could assist to identify the hard one whose clothing, pose, and makeup have changed or even failed to detect the face. For instance, in Figure 1, the man in red box is easily recognized with clear facial features, while the woman is hard to identify. However, if we realize they are a "couple," then it would be easy to get her identity. In addition, even if the man is hard to identify, we could still narrow down probable candidates into characters who formulate a couple relationship. The procedure is also consistent with the human way of thinking when identifying illegible characters from long videos. Nevertheless, it is non-trivial to only utilize visual content to reveal the social relation between characters in a particular scene. Therefore, in this article, we propose to leverage the emerging novel time-sync comments together, or so-called "bullet-screen comments" (crowd-sourced comments) [24], which could provide rich comprehensive context information in a plot. For example, in the right part of Figure 1, we list several comments that could benefit the revealing of the "couple" relationship.

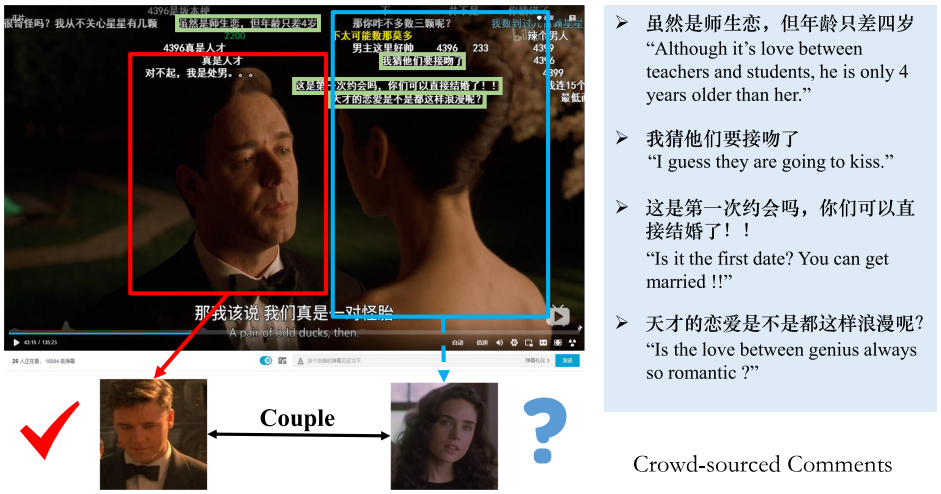


Fig. 1. Grounding textual cues (e.g., time-sync comments) within a video clip could perform as social context, thus provide social relation cues for person identification under occluded conditions.

To that end, in this article, we propose a relation-aware framework to deal with the person search task in a graph-based approach, which utilizes rich semantic contexts to profile characters and potentially assist person search in a way consistent with human thinking.

Specifically, we first segment videos into several semantically independent scene units and focus on each scene unit, respectively, to obtain more stable social contexts and limit the variation of visual information. Then, considering that the graph structure is flexible to abstract out the contextual information among various characters, we construct inner-personal links through a graph formulation operation for each scene, which connects query characters with candidate characters as graph nodes. Finally, we run a social context-aware label propagation to aggregate contextual information and identify those video characters. During the propagation, low-level cues (i.e., visual features) and high-level semantic cues (i.e., social relations) could benefit each other to get more accurate predictions in two aspects mentioned above: utilizing easily recognized characters to assist othersand narrowing down probable candidates. Our contributions could be summarized as follows:

(1) We study the problem of person search under complex conditions and utilize social relations with visual features complementarily to profile characters to support person search in a graph-based way.

(2) We utilize high-level multi-modal semantic context cues effectively by semi-supervised label propagation and combine both high-level semantic cues and visual cues for identification in a reasonable way.

(3) The extensive experiments on a real-world dataset prove the effectiveness of our method, especially when characters are hard to identify with visual cues.

2 RELATED WORK

In this section, we first summarize prior work on *person search* and then review some related techniques such as *label propagation* and *multimodal learning*.

2.1 Person Search

Person search aims at searching a target person in a series of image galleries without bounding box annotation, which can be seen as a combination of pedestrian detection [41, 42] and person

re-identification (re-ID) [8, 11, 23, 27]. Depending on whether the two modules are jointly designed, the person search framework can be roughly divided into two types. The first one focuses on building the framework with pedestrian detection and person re-ID separately [7, 23]. They first detect the person region and then match the person region pairs with person re-ID module. Another one usually utilizes multi-level factorization to map the visual features into latent discriminative factors at multiple semantic levels without bounding box annotation [6] and adapt jointly optimizing to learn the joint model [36]. These works are usually based on **convolutional neural networks (CNN)** and mostly aim at learning a non-linear mapping that transforms person regions into a common embedding space, in which the distances between queries and gallery images are learned [8, 11]. Some other works also utilize **residual networks (ResNet)** [16] or direct vectorization [28] to improve the performance. Recently, spatial matching and multi-scale features are added to gain the spatial information [2, 22, 43, 44], assigning different important weights to different regions, and further boost the performance of person search task.

However, the datasets that traditional person search methods use are captured by just several cameras in nearby locations within a short period, so the visual appearance of the same identities are usually similar, with few occlusion, clothing, and pose changes. Consequently, adopting conventional methods for person retrieval in movies directly is difficult to gain ideal performance.

2.2 Graphical Label Propagation

Label propagation [34, 37] is a widely used semi-supervised graph-based learning method that propagates the label attributes in a graph. When utilizing the label propagation for person retrieval task [18, 33, 45], the graph nodes represent the visual data and graph edges represent their pairwise similarities. Usually the initially labeled nodes are annotated query persons, and the rest unlabeled nodes are gallery images. Then, label inference is performed along graph paths that connect labeled nodes to unlabeled ones. A few kinds of label propagation methods have been proposed to tackle object detection and face recognition problems in computer vision field [21, 32]. They mostly focus on getting a more reliable node representation or adjusting the graph construction to adapt to the task scenario. In this article, we take the node relation, i.e., the person relation, into consideration. They form an invisible relation network between special nodes additionally, and also provide inferable cues in another semantic level to jointly resolve the difficulty of person search in dynamic scenarios.

2.3 Multi-modal Learning

With the quantity of digitized multimedia content increasing dramatically, multi-modal learning methods [1, 14] have been widely adapted to process and relate information from multiple modalities. Some of them focus on fusing visual and textual information, e.g., M-DBN [29] introduced a multi-modal deep belief network for each modality and combined them into joint representations. M-DBM [30] extended the M-DBN to multi-modal deep Boltzmann machines, allowing for the low-level representations of each modality to influence each other. Unified VSE [35] proposed a contrastive learning approach for the effective learning of alignment from image-caption pairs. These methods are usually based on well-annotated tasks such as image-caption and structured modality information, e.g., mapping the certain semantic component in sentences to the corresponding area of images. Hence, they are hard to fit with complex videos with variegated frames and tremendous unorganized textual documents.

Meanwhile, how to fuse visual modality and semantic context for person retrieval task is still under-explored. As for crowd-sourced comments in online videos, it could provide explicit semantic cues in user emotion analysis [20], but could be challenging for being used as the assistance of person retrieval, especially with semi-supervised graph-based learning method. Reasons are as

follows: First, contents of crowd-sourced comments within a time period are usually relevant to all appeared roles during the period, lacking directionality for a specific role. Second, crowd-sourced comments are usually high-level semantic and subjective comments for current plots and social scenarios, which are not binding description features of characters. With these problems, we try to mine semantic cues at a higher level, revealing social relations to adapt to these features of crowd-sourced comments, thus benefiting each other with visual information and assisting with identity inference in person search task.

3 TECHNICAL FRAMEWORK

In this section, we will formally define our problem with preliminaries and then introduce our framework in detail, including the design of modules step-by-step.

3.1 Preliminary and Problem Definition

As mentioned above, we target at the person search task in videos. For ease of description, we have $\boldsymbol{V} = \{f_{t_1}, f_{t_2}, \ldots, f_{t_n}\}$, a streaming collection of frames to denote a video, with f_t representing the frame at timestamp t. Each frame f_t may contain several **regions of interest (RoIs)** denoted as $\boldsymbol{R}_t$, in which each RoI $\in \boldsymbol{R}_t$ contains a specific person. Definitely, $\boldsymbol{R}_t = \emptyset$ indicates that no person appears in frame f_t. Note that these RoIs, i.e., person bounding boxes, are detected by the detection module, which will be discussed in Section 3.3.

At the same time, long videos from online social media platforms are often accompanied with numerous time-sync texts including subtitles and crowd-sourced comments. We use $\boldsymbol{D} = \{d_{t'_1}, d_{t'_2}, \ldots d_{t'_m}\}$ to denote textual documents, where $d_{t'}$ represents the text at timestamp t'.

Finally, in person search task, the query is denoted as $\boldsymbol{Q} = \{\boldsymbol{R}_q, G_r\}$. $\boldsymbol{R}_q = \{RoI_{q_1}, RoI_{q_2}, \ldots, RoI_{q_c}\}$ contains a small amount of manually labeled RoIs for each target character, in which q_i indicates a specific identity from C target characters. At the same time, relation graph G_r leads the way to utilize social context cues to benefit our person search task. In detail, G_r is formulated according to these target characters and their manually labeled social relationships. Along this line, if two target characters q_i and q_j form a certain relationship r_{ij}, there will be an edge labelled as r_{ij} connecting these two nodes representing q_i and q_j in G_r. In conclusion, the person search problem can be defined as follows:

Problem definition: Given the video $\boldsymbol{V}$ with corresponding textual documents $\boldsymbol{D}$, as well as query for target characters $\boldsymbol{R}_q$ formulated as graph G_r according to their social relationships. For each target character q_i in the query, we aim at revealing all the frames that contain q_i from $\boldsymbol{V}$.

3.2 Framework Overview

As mentioned above, our goal is to search person in a social context-aware condition. There are two main issues in this task. First, it is non-trivial to transform the label propagation into a relation-aware procedure. Second, how to reveal relationship cues from varying scenes in long videos also remains a problem. To deal with these problems, we propose a framework that contains three modules, as illustrated in Figure 2. Functions of these modules are introduced as follows:

(1) First, we utilize the **detection module** to locate characters and get unidentified RoIs, and then a projection is conducted to obtain these RoIs' visual embedding.

(2) Second, as a pre-processing step, we segment a video into a series of scene units, then the **graph formulation module** organizes these RoIs through obtained scene units to formulate graphs.

(3) Finally, the **propagation module** runs relation-aware label propagation over formulated graphs to predict the identity of each RoI.

Related technical details will be detailed in the following subsections.

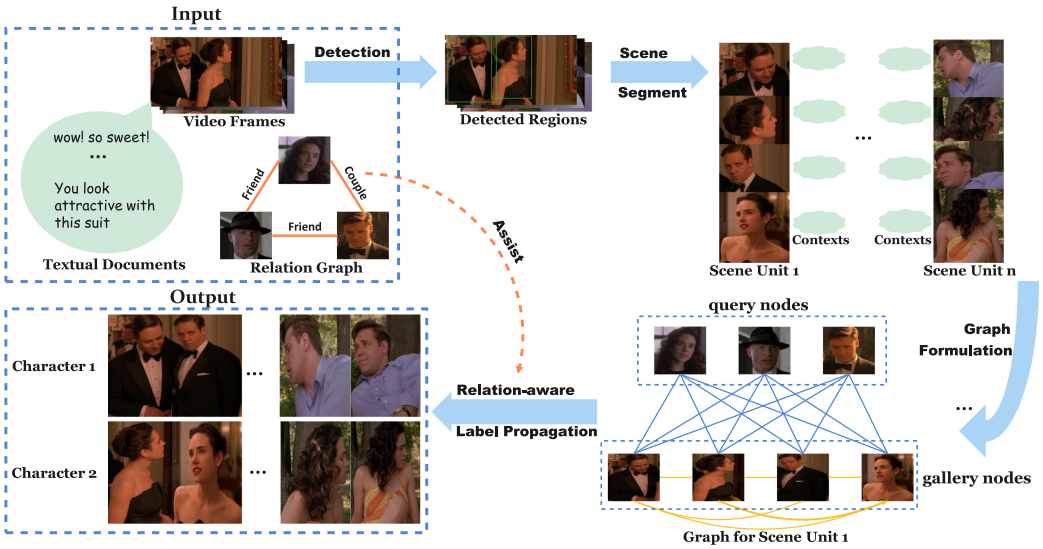


Fig. 2. Pipeline of overall framework. Given the input containing raw frames and textual documents as well as annotated relation graph between target characters, we first run a person detector to get RoIs. Then these RoIs are organized into several scene units according to the scene segments of videos, and textual documents within each scene units are treated as social contexts. The detected unidentified RoIs are connected through graph formulation (connections between queries and galleries are in blue, and connection between galleries are in yellow). Finally relation-aware label propagation is conducted over built graph with the assistance of relation graph to get the inference.

3.3 Detection Module

Given a set of frames $\boldsymbol{V}$, the detection module targets at capturing and embedding all potential RoIs for both target characters and non-target characters without discrimination, which can be seen as a pre-processing step. It is divided into two parts: capturing part and embedding part.

In the capturing part, we adapt some state-of-the-art person detectors to obtain the person regions $\{\boldsymbol{R}_{t_1}, \boldsymbol{R}_{t_2}, \ldots, \boldsymbol{R}_{t_n}\}$ from $\boldsymbol{V}$ without distinction. It is worth noting that the timestamp t_i of each RoI $\in \boldsymbol{R}_{t_i}$ is also obtained from its origin frame f_{t_i} and kept for the graph formulation module.

In the embedding part, we extract both body features and face features for obtained person regions, which is out of the consideration that body features and face features are complementary when searching person in long videos. To be specific, body features perform better in a short term, since they usually focus on the temporal attributes such as clothing and hair, which are relatively more stable than face emotions and poses in a short term. In contrast, queries are usually collected from a concrete timestamp that is far away from most candidates scattered over the timeline, thus varying a lot from candidates on body attributes. Face features with long-time immutability is suitable to link queries and candidates in such conditions. Hence, we take both body features and face features, utilizing several existing person re-identification models and face-recognition models to embed these features.

3.4 Graph Formulation Module

As mentioned before, semantic cues within movies are complex and mixed, implicit in different scenarios and hard to distinguish. To refine relation contexts to get more stable relation cues and limit the variation of visual information, we first segment RoIs into several short and semantically

independent scene units. Specifically, we adopt *PySceneDetect*,[1] a scene partition tool based on capturing scene changes and conducting advanced content-aware fast-cut detection for videos, to segment the videos into semantically independent short clips. The RoI set $\{\boldsymbol{R}_{t_1}, \boldsymbol{R}_{t_2}, \ldots, \boldsymbol{R}_{t_n}\}$ is also organized into several disjoint scene units following the timestamp of each RoI. Intuitively, RoIs whose timestamps are within the same clip will share a same scene unit. These organized RoIs could be represented as $\{\boldsymbol{R}_{s_1}, \boldsymbol{R}_{s_2}, \ldots, \boldsymbol{R}_{s_{N_c}}\}$, with N_c representing the number of video clips. Correspondingly, we organize subtitles and crowd-sourced comments into $\{d_{s_1}, d_{s_2}, \ldots, d_{s_{N_c}}\}$, with d_{s_i} containing all texts whose timestamps are within the ith scene unit. Through this way, RoIs as well as texts within the same scene unit are usually taken from the same scenario and share the same semantic context due to the situational similarity and closeness along the timeline.

Afterwards, the scenario graph construction is performed over every scene unit. For each scene unit s_i, we select $\boldsymbol{R}_{s_i} \cap \boldsymbol{R}_q$ as graph nodes to construct G_{s_i}, where all $g_i \in \boldsymbol{R}_{s_i}$ are marked as gallery nodes, and all $q_i \in \boldsymbol{R}_q$ are marked as query nodes. Any two of gallery nodes are linked by an edge, and there is also an edge between each query and gallery node pair. Specifically, G_{s_i} would contain $|\boldsymbol{R}_{s_i}| + |\boldsymbol{R}_q|$ nodes, $|\boldsymbol{R}_{s_i}|(|\boldsymbol{R}_{s_i}| - 1)/2$ edges between gallery nodes, and $|\boldsymbol{R}_q| \cdot |\boldsymbol{R}_{s_i}|$ edges between query and gallery nodes. In particular, the identities of all query RoIs are known, while the gallery nodes remain unknown.

3.5 Relation-aware Propagation Module

The Relation-aware propagation module targets at propagating identities from labeled nodes, i.e., $q_i \in \boldsymbol{R}_q$ to unlabeled nodes, i.e., $g_i \in \boldsymbol{R}_{s_k}$ through links between node pairs. We associate each node with a $|\boldsymbol{R}_q|$-dimensional probability vector p to denote its identity, as well as several aforementioned embedding to denote its feature. Initially, p_i is set to one-hot vector for each q_i that indicates the identity, while for all gallery nodes it is set to zero vector. The overall algorithm is presented in Algorithm 1.

3.5.1 Relation Detection. First, we combine social contexts from different modals, i.e., textual and visual information within the current scene unit, to reveal the social relationships implicit in video scenarios.

Assume g_i and g_j are two gallery nodes picked from the scene unit s_k. For textual information, we extract all texts within s_k as a document and get its *tf-idf* feature $\boldsymbol{e}_{s_k}$. Then, following the idea of Reference [31], we extract the gender features $\boldsymbol{e}_{ij}^{gen}$, as well as age features $\boldsymbol{e}_{ij}^{age}$ and activity features $\boldsymbol{e}_{ij}^{act}$ of g_i and g_j as the visual aspect of social contexts. A calibrated **support vector machine (SVM)** combining all features above is trained to infer the social relations[2]:

$$\boldsymbol{c}_{ij} = \left[\boldsymbol{e}_{s_k} \middle\| \boldsymbol{e}_{ij}^{gen} \middle\| \boldsymbol{e}_{ij}^{age} \middle\| \boldsymbol{e}_{ij}^{act}\right], \tag{1}$$

$$(r_{ij}, p_{ij}) = SVM(\boldsymbol{c}_{ij}), \tag{2}$$

where $\|$ denotes the concatenate operation, r_{ij} is the inferred relationship between g_i and g_j, and p_{ij} is the corresponding probability. Intuitively, the visual aspect of social contexts could performs as a fine-tuning and supplement to textual information. For instance, in the family scenario, father-son(kinship) and mother-father(couple) are two different relationships, but the textual information may share a same topic of family daily life, which brings ambiguities. However, by incorporating visual aspect of social contexts such as ages and genders, it is much easier to distinguish these two relationships.

[1] https://py.scenedetect.com.

[2] Details are presented in the Appendix.

ALGORITHM 1: Relation-aware Label Propagation

Input: Query nodes: $\{q_i\}_{i=1}^{C}$; Query label vector: $\{\boldsymbol{v}_i^q\}_{i=1}^{C}$;
Gallery nodes: $\{g_i\}_{i=1}^{L}$; Gallery label vector: $\{\boldsymbol{v}_i^g\}_{i=1}^{L}$;
Contextual information: $\{\boldsymbol{c}_{ij}\}$; Parameter α_r, K
Output: Updated label vector $\{v_i^g\}_{i=1}^{L}$

Initialize matrix S_{qg}, $\omega_{qg} \in \mathbb{R}^{L\times C}$
Initialize matrix S_{gg}, $\omega_{gg} \in \mathbb{R}^{L\times L}$
for i $\leftarrow$ 1 to L **do**
  for j $\leftarrow$ 1 to C **do**
    $S_{qg}[i][j] = S^r(\{\boldsymbol{c}_{ik}\}_{g_k\in\kappa(g_i)}, g_i, q_j, \alpha_r)$
  end for
  for j $\leftarrow$ 1 to L **do**
    $S_{gg}[i][j] = S^r(\boldsymbol{c}_{ij}, g_i, g_j, \alpha_r)$
  end for
end for
for i $\leftarrow$ 1 to L **do**
  for j $\leftarrow$ 1 to C **do**
    $\omega_{qg}[i][j] = \frac{S_{qg}[i][j]}{\sum_{k=1}^{C} S_{qg}[i][k]+\sum_{k=1}^{L} S_{gg}[i][k]}$
  end for
  for j $\leftarrow$ 1 to L **do**
    $\omega_{gg}[i][j] = \frac{S_{qg}[i][j]}{\sum_{k=1}^{C} S_{qg}[i][k]+\sum_{k=1}^{L} S_{gg}[i][k]}$
  end for
end for

Update strategy:
for *iter* $\leftarrow$ 1 to K **do**
  if *linear diffusion* **then**
    for i $\leftarrow$ 1 to L **do**
      $\boldsymbol{v}_i^g = \sum_{j=1}^{C} \omega_{qg}[i][j] * \boldsymbol{v}_j^q + \sum_{j=1}^{L} \omega_{gg}[i][j] * \boldsymbol{v}_j^g$
    end for
  else if *prudent update* **then**
    for i $\leftarrow$ 1 to L **do**
      for l $\leftarrow$ 1 to C **do**
        $\boldsymbol{v}_i^g[l] = max\{\underset{k\le C}{max}\{\omega_{qg}[i][k] * \boldsymbol{v}_j^q[l]\}, \underset{k\le L}{max}\{\omega_{gg}[i][k] * \boldsymbol{v}_j^g[l]\}\}$
      end for
    end for
  end if
end for

return $\{\boldsymbol{v}_1^g, \boldsymbol{v}_2^g, \ldots, \boldsymbol{v}_L^g\}$

3.5.2 Relation-enhanced Similarity Measure. To transform the propagation into a relation-aware procedure, we incorporate the relation graph G_r into the built scenario graph to get a more stable similarity measurement. Since movie scenarios are usually with high diversity, e.g., clothing and hair as well as gestures of characters would change frequently, it is difficult to obtain stable

visual cues, bringing errors during the propagation. At the same time, social contexts are general descriptions, which are relatively ambiguous, as different persons pairs could share similar social contexts.

Along this line, we propose to utilize a deeper fusion of visual and relation cues to complement each other to enhance the similarity measurement. We divide the the similarity between any node pairs n_i and n_j into two parts: visual similarity S^v and relation-guided similarity S^r. The S^v is measured through cosine similarity between visual embedding:

$$S^v(n_i, n_j) = \alpha_b \cdot \left\langle \boldsymbol{e}_i^{bo}, \boldsymbol{e}_j^{bo} \right\rangle + (1 - \alpha_b) \cdot \left\langle \boldsymbol{e}_i^{fa}, \boldsymbol{e}_j^{fa} \right\rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between two feature vectors, and $\boldsymbol{e}^{bo}$ as well as $\boldsymbol{e}^{fa}$ represent body features and face features, respectively, and α_b is used for adjusting the weight. Particularly, α_b is set to 1 when any of these two nodes failed to detect faces.

S^r performs as the deeper fusion of visual and relation cues and is measured conditioned on the type of nodes. We introduced two different measurement for Sim^r: (A) S^r between query nodes and gallery nodes, and (B) S^r between gallery nodes.

(A) S^r between query nodes and gallery nodes. To measure the relation-aware S^r between a query node q_i and a gallery node g_j, we first reveal the social relationships between g_j and each of its neighbors nodes $g_k \in \kappa(g_j)$ in galleries:

$$\mathbb{G}_j^r = \{(r_{jk}, p_{jk}) \mid g_k \in \kappa(g_j); (r_{jk}, p_{jk}) = SVM(\boldsymbol{c}_{jk})\}. \tag{4}$$

Then, query nodes whose represented characters are unlikely to co-occur with g_j in videos are filtered out. We keep those remains as well as corresponding relationships that could support their appearances:

$$\mathbb{Q}_j^r = \{(q_m, r_{jl}) \mid G_r(q_i, q_m) = r_{jl}; (r_{jl}, p_{jl}) \in \mathbb{G}_j^r\}, \tag{5}$$

where $G_r(\cdot, \cdot)$ returns the relation represented by the edge connecting two nodes in G_r. Finally, the strongest co-occurrence evidence is selected to support the measurement of $S^r(q_i, g_j)$:

$$S^r(q_i, g_j) = \max\{p_{jk} \cdot S^v(q_m, g_k) \mid (r_{jk}, p_{jk}) \in \mathbb{G}_j^r; (q_m, r_{jl}) \in \mathbb{Q}_j^r, r_{jk} = r_{jl}\}. \tag{6}$$

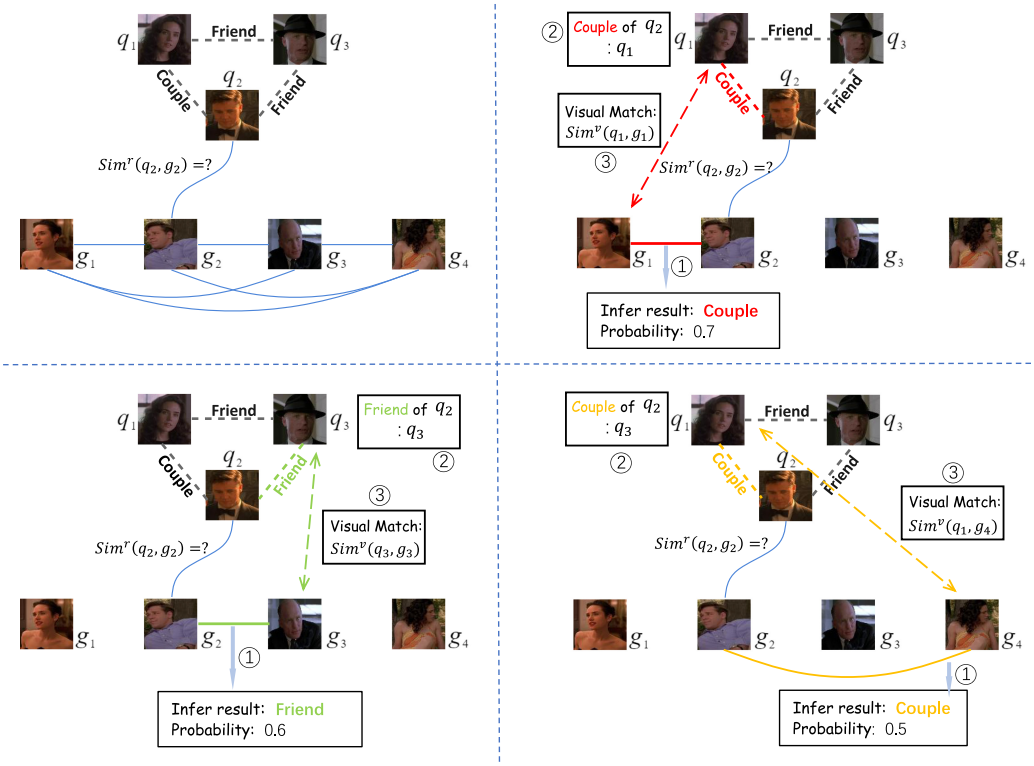
Through Equations (5) and (6), the stable cues for co-occurrence could be obtained via a relation-aware transformation. Figure 3 presents an instance for S^r between query nodes and gallery nodes. To measure $S^r(q_2, g_2)$, relationships between g_2 and its neighbor gallery nodes includes (g_1, g_3, and g_4), and relationships between g_1 and these three neighbors are inferred successively at the first step, which are *Couple*, *Friend*, and *Couple*. Then query nodes that form these relationships with q_2 are picked, including q_1, q_3, and q_1. Finally, visual match is performed between picked query nodes and corresponding neighbor gallery nodes that formulate the same relationships with q_2 and g_2, respectively, i.e., $S^v(q_1, g_1)$, $S^v(q_3, g_3)$, $S^v(q_1, g_3)$. The maximal match is picked as the Sim^r between q_2 and g_2.

(B) S^r between gallery nodes. As mentioned before, besides co-occurrence cues, associations between characters and scenarios could also be revealed by relationships. We incorporate such associations to enhance the measurement of S^r between gallery nodes. For any two gallery nodes g_i and g_i from the same scenario, their relationship is inferred first:

$$\mathbb{G}_i^r = \{(r_{ij}, p_{ij}) \mid (r_{ij}, p_{ij}) = SVM(\boldsymbol{c}_{ij})\}. \tag{7}$$

Taking relationships as associations between scenarios and characters, we filter out query nodes irrelevant to current scenarios and narrow down potential candidates guided by relationships:

$$\mathbb{Q}_{ij} = \{(q_k, q_l) \mid G_r(q_k, q_l) = r_{ij}; (r_{ij}, p_{ij}) \in \mathbb{G}_i^r\}. \tag{8}$$



$$Sim^r(q_2, g_2) = \max\{0.7 * Sim^v(q_1, g_1), 0.6 * Sim^v(q_3, g_3), 0.5 * Sim^v(q_1, g_4)\}$$

Fig. 3. Calculation of Sim^r between query nodes and gallery nodes. This figure illustrates the calculation of $Sim^r(q_2, g_2)$ between q_2 and g_2 as an example. We try to find all potential relation cues to support the $Sim^r(q_2, g_2)$, and the detailed process of each relation cue is presented in each subfigure. The order numbers in each subfigure denote the sequence of processing.

Note that (q_k, q_l) and (q_l, q_k) are treated as different elements in $\mathbb{Q}_{ij}$. To get the strongest support, i.e., most relevant characters to the current scenario, we select query node pairs that are most likely the mapping of corresponding gallery nodes as follows:

$$(q_{k'}, q_{l'}) = \underset{(q_k, q_l)}{argmax} \left\{ \min\{S^v(q_k, g_i), S^v(q_l, g_j)\} \,\middle|\, (q_k, q_l) \in \mathbb{Q}_{ij} \right\}. \tag{9}$$

With Equation (9), (g_i, g_j) are mapped to the most likely corresponding query pair $(q_{k'}, q_{l'})$ through associations between scenarios and characters, which is revealed by relationships. Finally, S^r between two gallery nodes is measured with similarity between the corresponded query pair:

$$S^r(g_i, g_j) = p_{ij} \cdot S^v(q_{k'}, q_{l'}). \tag{10}$$

The S^r transfers the similarity measurement between gallery nodes to a more reliable comparison between query nodes through the association between scenarios and characters, and an instance for the S^r between gallery nodes is illustrated in Figure 4. To get $S^r(g_2, g_3)$, relationship *Friend* between g_2 and g_3 is inferred first. Then all query node pairs that form *Friend* in G_r are picked, including (q_1, q_3), (q_3, q_1), (q_2, q_3), and (q_3, q_2). Visual matches between each gotten pair and (g_2, g_3) are performed, which are denoted as four different colors in the figure, with possibilities of 0.4, 0.5,

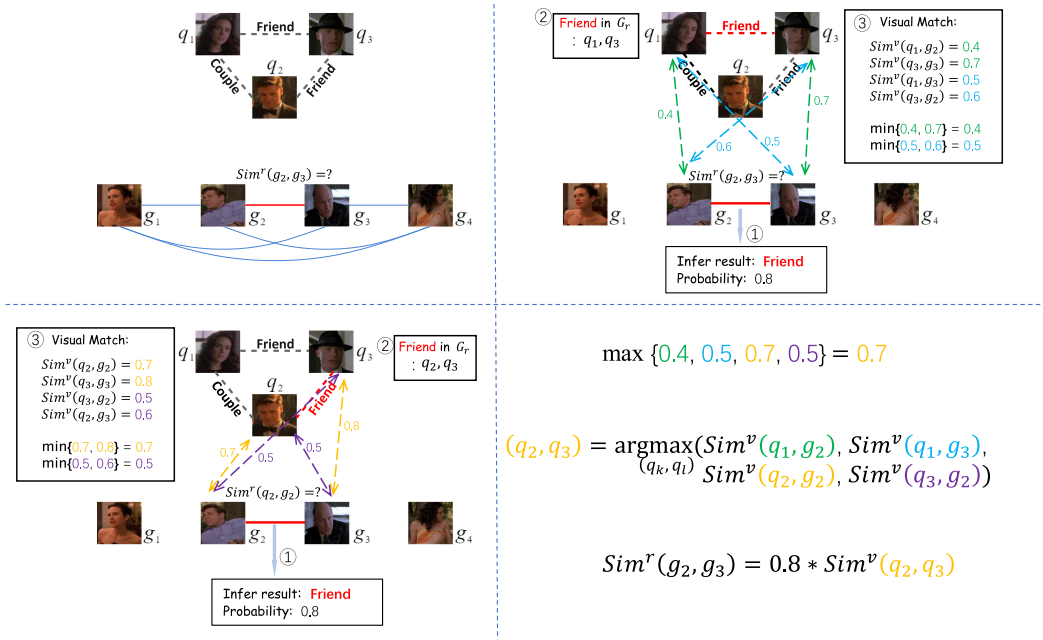


Fig. 4. Calculation of Sim^r between gallery nodes. This figure illustrates the calculation of $Sim^r(g_2, g_3)$ between g_2 and g_3 as an example. We try to find all potential relation cues according to the relation graph to a get more reliable measurement for the similarity between g_2 and g_3, and the detailed process of each relation cue is presented in each subfigure. The order numbers in each subfigure denote the sequence of processing.

0.7, and 0.5. Finally the most likely mapping, i.e., the match in yellow representing the mapping from (g_2, g_3) and (q_2, q_3) is selected. The $Sim^r(g_2, g_3)$ is measured as the match of (q_2, q_3).

With discussions above, for each node pair (n_i, n_j), we combine S^v and S^r as the final similarity measurement:

$$S(n_i, n_j) = S^v(n_i, n_j) + \alpha_r \cdot S^r(n_i, n_j), \tag{11}$$

where n_i and n_j could represent both query nodes and gallery nodes. Overall, S^r fuses relations and visual information in a complementary method, assisting visual match with relation cues and utilizing visual information to gain stronger and more reliable relation cues at the same time.

Through this way, we fuse visual information and semantic information in a higher semantic level through the link of social contexts. Another advantage is that it is compatible with different visual features, which could be replaced flexibly conditioned on the types of videos we use.

3.5.3 Update Strategy. Traditional update strategies of label propagation are usually based on linear diffusion, where each node would update its probability vector by taking a linear combination of those from neighbors. The scheme can be expressed as follows:

$$\boldsymbol{p}_i^{(t+1)} = \sum_{j \in \mathcal{N}(i)} \boldsymbol{\omega}_{ij} \boldsymbol{p}_j^{(t)}, \quad \text{with } \boldsymbol{\omega}_{ij} = \frac{S(n_i, n_j)}{\sum_{j' \in \mathcal{N}(i)} S(n_i, n_{j'})}. \tag{12}$$

The formula above presents the update process of probability vectors from iteration-t to iteration-$(t+1)$. Here, $\mathcal{N}(i)$ represents all the neighbor nodes of node i. Through this approach, label probabilities are propagated between each other, and nodes with high similarities tend to share the same identity finally.

Reference [17] proposed an enhanced update strategy to adapt to person search task, which is more prudent during the propagation and only trusts neighbors with strong evidence. The enhanced iteration strategy could deal with noisy data in person search datasets, and it can be expressed as follows:

$$\begin{aligned} \boldsymbol{p}_i^{(t+1)}[c] &= \max_{j \in \mathcal{N}(i)} \{\boldsymbol{\omega}_{ij} \boldsymbol{p}_j^{(t)}[c]\}, \\ \text{with } \boldsymbol{\omega}_{ij} &= \frac{S(n_i, n_j)}{\sum_{j' \in \mathcal{N}(i)} S(n_i, n_{j'})}. \end{aligned} \tag{13}$$

Here, $p[k]$ represents the value of vector p in the kth dimension. This scheme followed the idea of only collecting the strongest evidence from neighbors. Intuitively, an identity is strongly supported for node i if one of its neighbors assigns a high probability to it.

With contents illustrated above, the propagation over scenario graph to get p_k can be considered as a coordinate ascent method to solve Formula (14). Here, $\mathcal{H}$ is the entropy, and $z_{kj}[c]$ indicate whether the neighbor n_j is a trustable source for the class c for the node n_k. In linear diffusion $z_{kj}[c]$ is set to 1 for any k, j, and c, and in prudent update it is set to 1 only when n_j provide strongest evidence to n_k in the class c, i.e., satisfying Equation (15).

$$\max \sum_{c=1}^{C} \boldsymbol{p}_k[c] \sum_{j \in \mathcal{N}(k)} z_{kj}[c]\, \boldsymbol{p}_j[c]\, S(n_k, n_j) + \sum_{c=1}^{C} \mathcal{H}\left(\boldsymbol{p}_k[c]\right) \tag{14}$$

$$j = \underset{k \in \mathcal{N}(i)}{argmax} \{\boldsymbol{\omega}_{ik} \boldsymbol{p}_k^{(t)}[c]\}. \tag{15}$$

4 EXPERIMENTS

4.1 Dataset and Data Pre-processing

4.1.1 Data Preparation. We conducted experiments based on movies collected from BiliBili,[3] one of the largest video-sharing platforms in China. Each movie lasts for about 1.5 to 2.5 hours and contains about 4,000 crowd-sourced comments and 1,400 subtitles. We selected five main characters from each movie, on average, as target characters. Along this line, 70 movies lasting 129.3 hours with 390 target characters were collected. We took 50 movies with 265 target characters as training set, with the rest as test set. We annotated time slots for every character during which it appeared, and sampled frames with one per 1.5 s. Finally, about 3,000 frames were collected for each movie, and we totally got 228K frames for all movies, among which 120K frames contain target characters.[4]

4.1.2 Data Pre-processing. To construct a high-quality query set $\boldsymbol{R}_q$, for each target character q_i, we picked a bright and clear frame with frontal face and then cut the target region out manually as its query RoI_{q_i}. At the same time, considering that crowd-sourced comments are usually informal expressions that contain a lot of emoji or slang, we took some regular rules to filter crowd-sourced comments. For example, comments that are too short, e.g., less than three words, will be removed first, since the average length of crowd-sourced comments is about nine words, and extremely short comments are usually meaningless. Besides, those meaningless sentences such as messy code and time/date marker were filtered to ensure the quality of textual information.

To build the relation graph G_r for each movie, six experts professional in social theories were invited to give accurate relationship annotations for all selected character pairs. The annotated

[3] www.bilibili.com.

[4] The whole dataset will be published soon.

relationships including workmate, kinship, hostile, friend, and couple. We picked relationships between character pairs that are supported by the most experts. To further improve the reliability of the relationship annotations, we filtered out relationships supported by less than three experts and kept the remaining. Finally, we got 431 relationships between 288 target characters, with six relationship pairs for each movie, on average.

4.2 Experiment Settings

4.2.1 Implementation Details. In this section, we will detail the implementation of detection module, graph formulation module, and propagation module.

Detection Module. As mentioned in Section 3, the detection module is composed of the capturing step and embedding step, while the former aims to capture RoIs from frames and the latter generates an embedding vector for each RoI. In the capturing step, we adopted Cascade R-CNN [4] as the detector to detect human regions in frames. The Cascade R-CNN was initialized with the *ResNet-101* backbone and pre-trained on ImageNet [9] and then fine-tuned on VOC2007 [13]. All parameters were kept the same as origin in our detection settings, and we finally selected human regions with the confidence score greater than 0.85 as detected RoIs.

As for the embedding step, we used both Re-ID features and face features for visual embedding. The Re-ID feature was embedded with **Cross-Level Semantic Alignment (CLSA)** [7], which is capable of learning more discriminative identity features. To further validate the effectiveness of our social context-aware structure, we also tried **Multi-Level Factorisation Net (MLFN)** [6] to embed body features.

The CLSA model was initialized with *ResNet-50* backbone and fine-tuned with 6,120 well-selected clear *RoI*s collected from 200 characters in our training set to fit the visual style of movies. The initial learning rate was set to 0.001, while the λ_{cc} and λ_{clsa} were set to 1 and 1. We used the top three blocks to construct the semantic pyramid. As for MLFN, five blocks were stacked and four factor modules are aggregated within each block. We used the same training set as in CLSA, with initial learning rate set to 0.001 and the fusion of deep feature and factor signature taken as visual embedding.

Correspondingly, faces were detected and embedded with MTCNN [39] and FaceNet [26]. We also took MTCNN [39] and ArcFace [10] to detect and embed faces to further validate the effectiveness. In MTCNN+FaceNet, faces were first aligned by MTCNN and resized to 160*160 to get face embedding, while in MTCNN+ArcFace faces were detected by MTCNN and resized to 112*112 and finally embedded by ArcFace. MTCNN+FaceNet was trained on VGGFace2 [5] and MTCNN+ArcFace was trained on MS1M-ArcFace [15]. Note that faces could be detected only in a few RoIs due to the diverse pose in movie scenario, with remaining RoIs failing to get faces and face embedding.

Graph Formulation Module. In the graph formulation module, We utilized the *PySceneDetect* to cut video into clips and record time intervals of each. The detection threshold was set to 40 while other parameters remain default. After the video segmentation, each video was partitioned into about 400 scenes on average with each scene lasting about 17.7 seconds, on average.

Propagation Module. We took social context-aware propagation method, in which the parameter α_r was set to 1.20 and will be discussed in parameter sensitivity experiments. As for the visual similarity, we combined the face similarity and body similarity for matching, with α_b set to 0.2 as in Reference [17]. The iteration time K was set to 20, and at each iteration in the prudent update, we followed the idea of fixing the labels of 10% nodes that have the highest confidence in [17], where the confidence was defined to be the maximum probability value in its class vector.

As for relation detection, age features and gender features were embedded by Rude Carnie [19]. We kept the default parameter setting as in the released model trained on Adience [12]. The feature in $fc7$ layer was taken as gender features and age features. For activity features, We used the published CNN-CRF model trained on SituNet [38]. The feature of the $fc7$ layer was extracted as activity features. Then, we concatenated these features as the visual feature to train the relation classifier.

4.2.2 Baseline Methods. We set up two groups of experiments, which adopted the linear diffusion and prudent update as update strategies during label propagation, respectively. In each group, we set up several baseline methods for comparison, which are illustrated as follows. We split 10% of the training set for tuning hyper-parameters.

Visual match-based approaches. which adopted traditional neural network methods including person re-identification and face recognition. We selected CLSA [7], MLFN [6] as person re-identification methods, and MTCNN [39]+FaceNet [26], MTCNN [39]+ArcFace [10] as face recognition methods. All baselines were either initialized with published models (if exist) and then fine-tuned on our movie dataset or trained from scratch on our movie dataset. The training parameters were kept the same as reported in original papers.

Visual context-aware approaches. which organized RoIs into several scene units and conducted traditional label propagation, aggregating visual contexts from neighbors of each node to get predictions. We tried label propagation with *MLFN* features and *FaceNet* features (LPMF), and label propagation with *MLFN* features and *ArcFace* features (LPMA), and label propagation with *CLSA* features and *FaceNet* features (LPCF), and label propagation with *CLSA* features and *ArcFace* features (LPCA). We used the same propagation graph as illustrated in Section 3.4 for fair comparison. The similarity between nodes was measured by weighted combination of the Re-ID similarity and the face similarity, with weights of 0.2 and 0.8, respectively, as in Reference [17] as our framework for fair comparison.

Visual & textual context-aware approaches. Meanwhile, we evaluate the performance of multi-modal methods combining textual features and visual features in label propagation as multi-modal baselines. Specifically, we collect texts for each RoI around its time windows with window size set to 5 s as documents to get its tf-idf feature as the textual feature. Then, we combine visual similarity and textual similarity as the similarity measurement, with the weight of textual similarity and visual similarity set to 0.20 and 0.80, respectively. These corresponding methods are denoted as TPMF (Visual & textual context-aware Propagation with *MLFN* features and *FaceNet* features), TPMA, TPCF, and TPCA. Note that there showed no significant correlation between the performance and the weight of textual features in our experiments, with more than one choice corresponding to the best performance. Finally, we reported the best performance on test set among these choices.

Social context-aware approaches. Similarly, we denote proposed Social Context-aware Person Search methods with different features as SCPS-MF (Social Context-aware Person Search methods with *MLFN* features and *FaceNet* features), SCPS-MA, SCPS-CF, and SCPS-CA.

4.3 Overall performance

In this section, we validated the overall performance of our framework. As mentioned before, the person search task aims at capturing all frames containing target characters. Thus, we selected Precision, Recall, and F1-value as evaluation metrics. The result is presented in Table 1 and Table 2, with linear diffusion and prudent update as update strategy, respectively. The face feature-based methods (FaceNet and ArcFace) outperformed body feature-based methods (MLFN and CLSA) significantly, for face features are much stabler in complex movie scenarios. Note that two face-based

Table 1. Overall Performance with Linear Diffusion

Backbone	Methods	P(%)	R(%)	F1(%)
visual match-based	MLFN [6]	31.18	39.55	34.87
	CLSA [7]	34.09	43.43	38.20
	MTCNN [39]+FaceNet [26]	57.76	41.95	48.60
	MTCNN [39]+ArcFace [10]	**62.98**	43.96	51.78
visual context-aware	LPMF [6, 17, 26]	45.75	60.63	52.15
	LPMA [6, 10, 17, 39]	46.38	61.81	52.99
	LPCF [7, 17, 26]	46.26	61.54	52.82
	LPCA [7, 10, 17, 39]	48.00	62.27	54.21
visual & textual context-aware	TPMF	41.13	48.34	44.44
	TPMA	42.07	49.15	45.34
	TPCF	42.15	49.00	45.31
	TPCA	43.50	51.01	46.96
social context-aware	SCPS-MF	53.41	67.13	59.49
	SCPS-MA	53.94	67.98	60.15
	SCPS-CF	54.09	67.91	60.21
	SCPS-CA	54.62	**68.54**	**60.79**

Table 2. Overall Performance with Prudent Update

Backbone	Methods	P(%)	R(%)	F1(%)
visual match-based	MLFN [6]	31.18	39.55	34.87
	CLSA [7]	34.09	43.43	38.20
	MTCNN [39]+FaceNet [26]	57.76	41.95	48.60
	MTCNN [39]+ArcFace [10]	**62.98**	43.96	51.78
visual context-aware	LPMF [6, 17, 26]	44.02	55.75	49.19
	LPMA [6, 10, 17, 39]	44.60	56.27	49.76
	LPCF [7, 17, 26]	44.39	56.21	49.60
	LPCA [7, 10, 17, 39]	45.97	56.86	50.84
visual & textual context-aware	TPMF	42.95	48.88	45.72
	TPMA	43.73	49.40	46.39
	TPCF	43.91	49.36	46.48
	TPCA	44.64	51.17	47.68
social context-aware	SCPS-MF	53.69	62.04	57.56
	SCPS-MA	54.19	62.25	57.94
	SCPS-CF	54.20	62.12	57.89
	SCPS-CA	54.75	**62.50**	**58.37**

methods could get high accuracy, for they only gave inferences to RoIs who had detected faces, with remaining RoIs treated as negative samples. As a result, they got poor performance in the recall rate. The graph-based (context-aware) methods achieved better performance compared to match-based methods, an important reason for which is that they conducted scenario partition and consider person search in each scenario unit, which simplified the complexities of scenarios.

Meanwhile, comparison between visual context-aware methods and social context-aware methods showed the significant improvement brought by relation cues, with about 6 to 7% in F1 value.

Table 3. Top Accuracy of Models with Linear Diffusion among All RoIs

Backbone	Methods	Top-1(%)	Top-3(%)	Top-5(%)
visual match-based	MLFN [6]	31.18	59.85	67.64
	CLSA [7]	34.09	62.97	69.40
	MTCNN [39]+FaceNet [26]	39.47	53.04	61.15
	MTCNN [39]+ArcFace [10]	41.9	54.72	61.35
visual context-aware	LPMF [6, 17, 26]	45.75	69.57	75.49
	LPMA [6, 10, 17, 39]	46.38	69.95	75.58
	LPCF [7, 17, 26]	46.26	69.93	75.57
	LPCA [7, 10, 17, 39]	48.00	70.40	76.27
visual & textual context-aware	TPMF	41.13	65.80	73.79
	TPMA	42.07	67.05	73.90
	TPCF	42.15	67.56	73.95
	TPCA	43.50	68.20	74.51
social context-aware	SCPS-MF	53.41	72.05	78.69
	SCPS-MA	53.94	72.87	79.00
	SCPS-CF	54.09	72.91	79.03
	SCPS-CA	**54.62**	**73.22**	**79.16**

Such improvement appeared in all SCPS methods with four kinds of features compared to corresponding baselines across both the linear diffusion and prudent update strategies.

The way of fusing multi-modal features played an important role in label propagation, with the simple combination of textual and visual features performing even worse than visual-based frameworks. An important reason is that crowd-sourced comments are usually high-level semantic comments made by multiple users and lack direct association with target characters.

Finally, the linear diffusion strategy performed better than the prudent update strategy. An important reason was that the dataset where prudent update strategy took was structured, with *RoI*s who share the same identity organized into a tracklet manually. In this case, more prudent strategies were required, since a wrong prediction would mislabel all RoIs along this tracklet. However, in more complex datasets such as unstructured movie frames, to identify a *RoI*, more visual contexts and semantic contexts were needed, so linear diffusion was able to gain a better performance.

We also conducted the significant test on performance, in which the best model of each group, i.e., LPCA, TPCA, and MTCNN+ArcFace, in linear diffusion were chosen as baselines. We used SCPS-CA as our validation model. Here, we propose a null null hypothesis, which is SCPS-CA outperformed to baselines with at least 7%, 6%, and 9% improvement in F1 value respectively. We conducted 10-fold cross-validation and get 10 result samples for each method. The significant testing was performed using the paired student's t-test with $p < 0.05$, which validated the effectiveness of our method.

We further validated the *top*-1, *top*-3, and *top*-5 accuracy of these methods. Note that to evaluate both the ability of extracting target characters from all and assigning correct label among RoIs of target characters, we set two groups of experiments. In the first group, experiments were conducted on all RoIs including numerous irrelevant characters, thus evaluating the ability of person search among data containing both target characters and irrelevant characters. In the second group, experiments were conducted on RoIs with irrelevant characters filtered, which evaluated the ability of assigning correct label among RoIs of target characters. The validation of frameworks with linear diffusion in two groups is listed in Table 3 and Table 4, while the framework with prudent update in two groups is listed in Table 5 and Table 6. Note that we took RoIs that failed to detect faces as false examples when evaluating the top accuracy performance of FaceNet and

Table 4. Top Accuracy of Models with Linear Diffusion among Target RoIs

Backbone	Methods	Top-1(%)	Top-3(%)	Top-5(%)
visual match-based	MLFN [6]	40.06	82.16	95.57
	CLSA [7]	43.48	82.74	95.69
	MTCNN [39]+FaceNet [26]	41.34	64.99	69.30
	MTCNN [39]+ArcFace [10]	45.72	65.21	69.33
visual context-aware	LPMF [6, 17, 26]	59.14	88.28	97.41
	LPMA [6, 10, 17, 39]	61.45	88.58	97.55
	LPCF [7, 17, 26]	59.92	88.39	97.51
	LPCA [7, 10, 17, 39]	62.72	89.05	97.99
visual & textual context-aware	TPMF	51.05	86.22	96.52
	TPMA	52.75	86.74	96.55
	TPCF	53.82	86.90	96.59
	TPCA	56.77	87.82	97.01
social context-aware	SCPS-MF	67.49	90.83	98.37
	SCPS-MA	67.93	91.61	98.46
	SCPS-CF	68.10	91.76	98.60
	SCPS-CA	**68.95**	**91.97**	**98.65**

Table 5. Top Accuracy of Models with Prudent Update among All RoIs

Backbone	Methods	Top-1(%)	Top-3(%)	Top-5(%)
visual match-based	MLFN [6]	31.18	59.85	67.64
	CLSA [7]	34.09	62.97	69.40
	MTCNN [39]+FaceNet [26]	39.47	53.04	61.15
	MTCNN [39]+ArcFace [10]	41.9	54.72	61.35
visual context-aware	LPMF [6, 17, 26]	44.02	70.44	77.80
	LPMA [6, 10, 17, 39]	44.60	70.75	77.83
	LPCF [7, 17, 26]	44.39	70.63	78.42
	LPCA [7, 10, 17, 39]	45.97	70.88	78.50
visual & textual context-aware	TPMF	42.95	70.21	77.69
	TPMA	43.73	70.35	77.76
	TPCF	43.91	70.40	77.78
	TPCA	44.64	70.75	77.84
social context-aware	SCPS-MF	53.73	72.66	78.98
	SCPS-MA	54.19	73.02	79.12
	SCPS-CF	54.20	73.11	**79.17**
	SCPS-CA	**54.75**	**73.13**	79.15

MTCNN+ArcFace. According to the result, we could observe similar improvement as in overall performance: Compared with several baselines, SCPS methods could achieve better performance in both two groups.

4.4 Ablation Study

At the same time, we turned to verify the effectiveness of different features in relation detection for the social context-aware person search. In this experiment setting, we used *CLSA* features as body features and *ArcFace* as face features.

Table 6. Top Accuracy of Models with Prudent Update among Target RoIs

Backbone	Methods	Top-1(%)	Top-3(%)	Top-5(%)
visual match-based	MLFN [6]	40.06	82.16	95.57
	CLSA [7]	43.48	82.74	95.69
	MTCNN [39]+FaceNet [26]	41.34	64.99	69.30
	MTCNN [39]+ArcFace [10]	45.72	65.21	69.33
visual context-aware	LPMF [6, 17, 26]	54.93	88.23	97.97
	LPMA [6, 10, 17, 39]	55.52	88.54	97.50
	LPCF [7, 17, 26]	55.29	88.27	98.01
	LPCA [7, 10, 17, 39]	56.50	88.41	98.14
visual & textual context-aware	TPMF	53.74	87.70	97.22
	TPMA	53.83	87.79	97.25
	TPCF	54.31	88.08	97.31
	TPCA	55.07	88.15	97.40
social context-aware	SCPS-MF	64.39	90.41	98.59
	SCPS-MA	64.87	90.96	98.65
	SCPS-CF	65.01	91.05	98.65
	SCPS-CA	**66.15**	**91.11**	**98.67**

Table 7. Ablation Study with Linear Diffusion

Methods	P(%)	R(%)	F1(%)
$SCPS_{w2c}$ w/o V	47.68	63.77	54.56
$SCPS_{tf-idf}$ w/o V	52.20	65.95	58.27
SCPS w/o T	44.77	59.23	50.99
$SCPS_{w2c}$	49.94	64.36	56.24
$SCPS_{tf-idf}$	54.62	68.54	60.79

Experimental results are listed in Tables 7 and 8, in which $SCPS_{w2c}$ w/o V and $SCPS_{tf-idf}$ w/o V represent SCPS methods whose relation detection tool uses *Word2vec* textual features only and *tf-idf* textual features only, respectively. Note that to get *Word2vec* features, we use the average of all word vectors to embed the corresponding document. SCPS w/o T represents SCPS method whose relation detection tool uses visual features only, i.e., age, gender features, and activity features. SCPS represents the complete SCPS method, in which the $SCPS_{tf-idf}$ represents our final framework, and $SCPS_{w2c}$ replaces the *tf-idf* features to *Word2vec* features.

Obviously, we could find that textual contexts were significant for the final performance, for SCPS achieved greater improvement compared with SCPS w/o T, with at least 5% and up to 9% in F1 value. Meanwhile, how to utilize these textual contexts also matters, with *Word2vec* feature-based methods getting about 4% worse performance than *tf-idf* feature-based methods. An important reason is that crowd-sourced comments contain large amounts of irrelevant emotional descriptions, which would bring a lot of noise when using the average of *Word2vec* vectors to model textual features. At the same time, visual features could act as additional parts to social contexts, for SCPS methods were about 2% better than its corresponding SCPS w/o V methods.

4.5 Parameter Sensitivity

Afterwards, we turned to measure the parameter sensitivity during the propagation, i.e., the sensitivity of the parameter α_r, which represents the weight of S^r. The result is shown in Figure 5, which

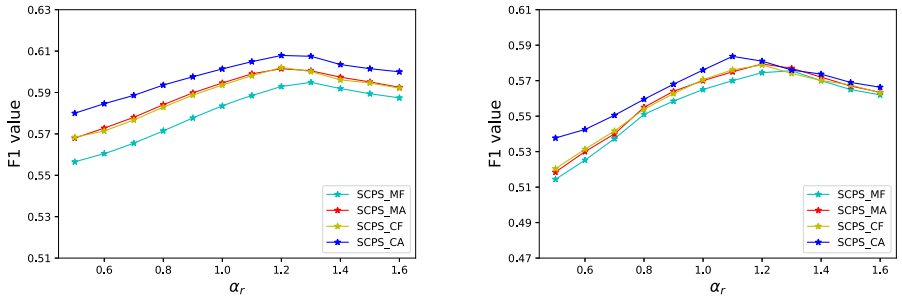
Table 8. Ablation Study with Prudent Update

Methods	P(%)	R(%)	F1(%)
SCPS_{w2c} w/o V	47.90	57.82	52.39
SCPS_{tf-idf} w/o V	51.93	61.03	56.11
SCPS w/o T	45.38	56.17	50.20
SCPS_{w2c}	50.27	59.45	54.48
SCPS_{tf-idf}	54.75	62.50	58.37

Table 9. Inference Latency of Different Kinds of Methods

Methods	per batch (in seconds)	per video (in hours)
CLSA	1.72	1.43
LPCA	3.16	2.65
SCPS-CA	3.38	2.81

We take CLSA as visual match-based methods, and LPCA as well as SCPS-CA as visual context-aware methods and social context-aware methods, respectively. The update strategy for context-aware methods is set to linear diffusion, and batch size is set to 32.



(a) Parameter sensitivity in linear diffusion. (b) Parameter sensitivity in prudent update strategy.

Fig. 5. The performance of social context-aware methods with different value of α_r.

presents the performance of SCPS methods with linear diffusion and prudent update strategy. As shown in the figure, with the increase of α_r, the overall performance of these social context-aware methods also grew in some degrees even when $\alpha_r > 1.0$. This phenomenon indicates that S^r could even play a more important role than visual information when the social contexts are rich and of high quality. At the same time, it could also inspire us to adjust α_r to a smaller value if the social contexts are not rich enough.

4.6 Inference Latency

We evaluated the time cost of our proposed method on a single Tesla K80 GPU. Table 9 shows the average latency in identifying a single batch of frames and getting appearance clips for all target characters in a movie. The context-aware methods were 1.84 times slower than match-based methods, and the extra times cost mainly lay in graph operations. However, compared to visual context-aware methods, our proposed social context-aware method brings only 7% extra time cost, since all social relation predictions as well as similarity weights between nodes could be pre-computed and cached.

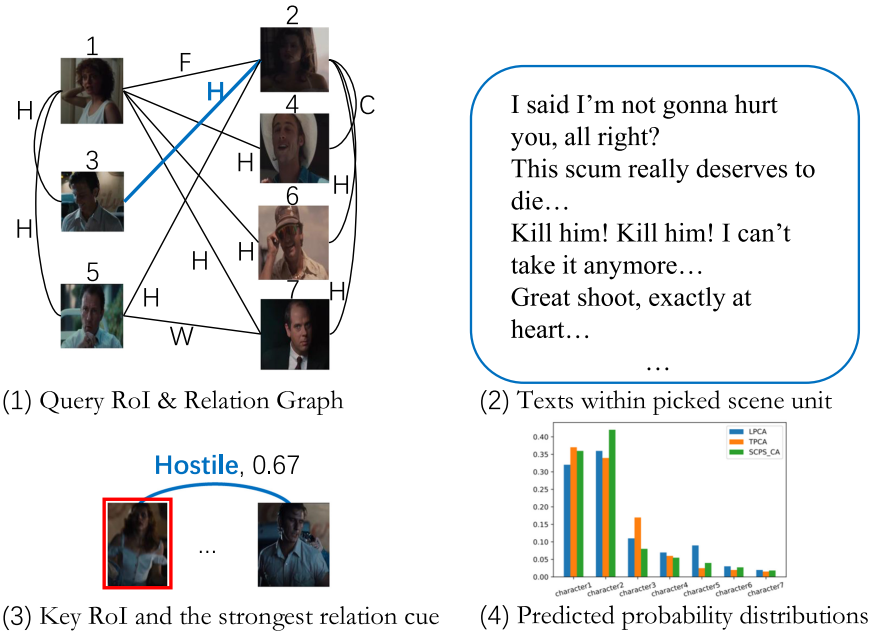


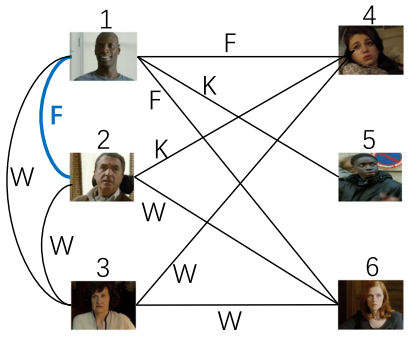
Fig. 6. A scene where characters are conflicting with each other. LPCA and SCPS-CA predict the key RoI correctly, and TPCA mislabels it as character 1.

Most importantly, the number of characters within a video is limited, which enables the offline computing of retrieving appearance clips for all potential target characters. In practical use, the online systems only need to fetch cached results that have been pre-computed by offline person search system as the response to users. In such situation, the performance is more worthy of people's attention in trading off the impact of efficiency degradation and performance improvement.

4.7 Case Studies

In this section, we present some examples to validate the effectiveness of social relation cues, as shown in Figures 6–8. Each figure is composed of four parts, as numbered at the bottom of each part. Part (1) contains query RoIs $\boldsymbol{R}_q$ and the relation graph G_r from a selected movie, in which we use K to denote *Kinship*, and W to denote *Workmate*, and F to denote *Friend*, and H to denote *Hostile*, and C to denote *Couple* for ease of description. Part (2) and part (3) contain two RoIs picked from a selected scene unit, as well as the corresponding textual contexts within the picked scene unit. In part (3), the RoI in red box is the key RoI that we focus on, and the rest one is the strongest social context cue supporting the prediction of the key RoI. We also present its predicted relation in part (3). In part (4) we show predictions of the key RoI from different models to illustrate the performance comparison between these models; here, we select LPCA, TPCA, SCPS-CA with linear diffusion as validation models.

Figure 6 is a case selected from the movie *Thelma and Louise*. We picked a scene where character 2 (the left RoI in part (3)) and character 3 (the right RoI in part (3)) are in conflict with each other. The key RoI (whose truth label is character 2) in red box is similar to the query RoI numbered as character 1 in hair cut and cloth, but varies from its corresponding truth query RoI in cloth. Through the result in part (4), we can observe that LPCA (in blue) predicts it correctly, but could not distinguish it with character 1 well, with probabilities in character 1 and character 2 are close.

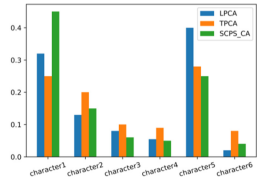


(1) Query RoI & Relation Graph

Nice man, getting a massage without forgetting his friend…
Is this the true friendship between men? …
Uhhhh, I also did this with my classmate last week…
…

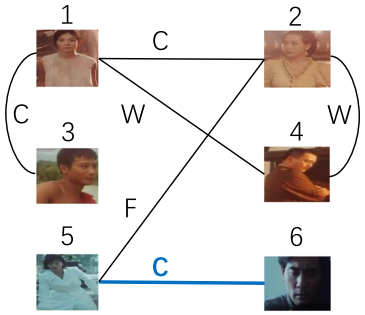
(2) Texts within picked scene unit

(3) Key RoI and the strongest relation cue



(4) Predicted probability distributions

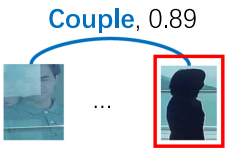
Fig. 7. Appeared characters are getting a massage in the selected scene unit. LPCA mislabels the key RoI as character 5, and TPCA gives better probability distribution, and SCPS-CA predicts it correctly.



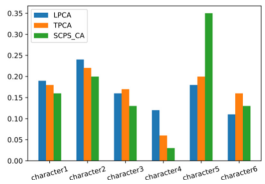
(1) Query RoI & Relation Graph

You don't want to get married?…
You don't love me anymore? …
Don't be silly, if you want, just get married…
Sorry, I fell in love with someone else…
…

(2) Texts within picked scene unit



(3) Key RoI and the strongest relation cue



(4) Predicted probability distributions

Fig. 8. A scene of lovers embracing while the key RoI is totally unrecognized. LPCA fails to distinguish it with other characters, and TPCA gives worse probability distribution, and SCPS-CA predicts it correctly.

Textual-incorporated TPCA (in yellow) mislabeled the key RoI as character 1. An important reason is that in this movie, character 1 and character 2 usually co-occur in most scenes, resulting in a high similarity between the textual contexts of them. This misguides the prediction of TPCA when probabilities given by visual-based approaches are close enough. SCPS-CA (in green) predicted

correctly by revealing the relation cue, i.e., *Hostile* formed by the key RoI and another presented RoI (whose truth label is character 3) in the picked scene unit, with another presented RoI that could easily be recognized as character 3.

Figure 7 is a case selected from the movie *Untouchables*, where we pick a scene in which the leading character (the left RoI, whose ground truth is character 1) is getting a massage with his friend (the right RoI, whose ground truth is character 2). The key RoI varies from its corresponding query RoI (character 1) in pose, clothing, with face incomplete and blocked by someone's hands. LPCA failed to distinguish it from the query RoI numbered as 5 due to the similar pose and color. TPCA lowered its predicted possibility in character 5 according to nearby textual contexts, for character 5 only appeared along with character 1 in scenarios where they interact as *Kinship*. SCPS-CA revealed the strongest relation cue in the picked scene unit, i.e., *Friend* relation with another RoI whose ground truth is character 2, and predicted it correctly enhanced by the relation cue.

Figure 8 is a case selected from the movie *Intimates*, where the picked scene is about a couple hugging in front of the window. The key RoI in red box whose truth label is character 5, is totally indistinguishable by visual-based approaches, for the face and cloth are all missing due to the ray of light, with only the silhouette remaining. TPCA got a worse prediction, which allocates the probability distribution even more uniformly, because there are more than one couple in this movie and textual contexts are highly similar between those characters. SCPS-CA predicted correctly by revealing the relation cue, i.e., *Couple* formed by the key RoI and another RoI in the picked scene unit whose truth label is character 6 and easy to recognize.

4.8 Discussion on Accumulative Errors

The proposed framework is relatively complicated with several modules, which may bring accumulative errors. Here, we provide several discussions about potential accumulative errors.

The detection module, which crops person areas in raw videos frames, and the relation detection module, which predicts relationship between graph nodes, are two potential modules most likely to bring accumulative errors. The published Cascade RCNN, which we took as the detection module, was evaluated on the VOC2007 and got 75.98% in AP50, limiting the error to a relatively low level. Meanwhile, as an integrated task, person search usually takes the person detection framework as pre-step inevitably and then conducts re-identification on obtained person regions. As revealed in existing person search works [7, 36], there is no strong evidence showing that joint modeling the detection module with Re-ID module performs better than modeling them separately. So, in this work, considering the computing resource, we modeled them separately.

As for the relation detection module, we kept predictions with confidence scores higher than 0.5 and got an accuracy of 73.19%, which is able to provide relatively high-quality relationship predictions for propagation section. Moreover, the weight of relation-aware similarities, i.e., α_r, could be adjusted in application depending on the quality of social context information. As shown in Section 4.5, since huge amounts of crowd-sourced comments are considered in this work, the best choice of α_r could reach 1.2, which is even higher than the weight of the original visual measurement. However, if only social contexts with lower quality are available, then we could reduce the weight α_r correspondingly.

5 CONCLUSION

In this article, we studied the person search task in movies. To better utilize rich semantic contexts in movies and mine its potential for profiling persons with high-level semantic cues, we proposed a graph-based framework, which incorporated social relation contexts to enhance the performance. To be specific, we first segmented videos into several independent scene units and located

characters indiscriminately through detection methods. Then, propagation graphs were built over detected RoIs within each scene unit and query RoIs. Finally, we conducted social context-aware label propagation, revealing implicit social relation cues within scene units, and then combined the relation graph and the built scene graph through social relation cues to give a more stable prediction. Experiments on plentiful movie datasets proved that our solution outperformed several state-of-the-art baselines and strongly validated the effectiveness of social contexts in the person search task.

APPENDICES

A DETAILS FOR RELATION DETECTION

A.1 Training data

We conduct the relation classifier training procedure based on the collected BiliBili dataset. For each selected character q_i, we have its relationships connected with others in G_r, as well as all frame-by-frame labeled time slots during which it appears, denoted as $\mathbb{T}_i = \{[t^i_{s_1}, t^i_{e_1}], [t^i_{s_2}, t^i_{e_2}], \ldots, [t^i_{s_{n_i}}, t^i_{e_{n_i}}]\}$.

Considering that characters associated with relationships may co-occur or appeared alternately in videos, for each edge $(q_i, r_{ij}, q_j) \in G_r$, we first aggregate time slots that are close enough:

$$\mathbb{T}^{'}_i = \left\{ \left[t^i_{s_k}, t^i_{e_{k+1}}\right] \,\middle|\, \left[t^i_{s_k}, t^i_{e_k}\right], \left[t^i_{s_{k+1}}, t^i_{e_{k+1}}\right] \in \mathbb{T}_i,\ t^i_{s_{k+1}} - t^i_{e_k} \le \tau \right\}, \tag{16}$$

where τ is set to 4.5 s in our settings. Then, we get all time slots during which q_i and q_j co-occurred or appeared alternately as follows:

$$\Omega_{ij} = \left\{ \left[t^i_{s_k}, t^i_{e_k}\right] \cap \left[t^j_{s_l}, t^j_{e_l}\right] \,\middle|\, \left[t^i_{s_k}, t^i_{e_k}\right] \in \mathbb{T}^{'}_i, \left[t^j_{s_l}, t^j_{e_l}\right] \in \mathbb{T}^{'}_j \right\}. \tag{17}$$

Then, we aggregate neared time slots in in Ω_{ij} until their spans reach 17 s, i.e., the average length of scene units as we observed in experiments. For each gotten time slot $[t_k, t_l]$, we collect all crowdsourced comments whose timestamps are within $[t_k, t_l]$ as textual features $\boldsymbol{t}_{kl}$, and corresponding gender, age, and activity features of q_i and q_j as visual features $\boldsymbol{v}_{ij}$. We take $(\boldsymbol{t}_{kl}, \boldsymbol{v}_{ij}, r_{ij})$ as a single training instance.

A.2 Classifier performance

We only collect instances on 50 movies of the training set to avoid the data leakage, where 70% of instances are used for training, 15% for validation, and the rest for testing. To obtain relation cues with higher accuracies, we only kept predictions with confidence score higher than 0.5, through which the classifier achieved an accuracy of 73.17% on the test set (among predicted instances).

B ANNOTATIONS FOR RELATIONSHIPS

B.1 Annotation procedure

We referred to the mainstream view in social domain theories [3, 25], which partitions social life into five social domains: attachment domain, reciprocity domain, mating domain, hierarchical power domain, and coalitional groups domain.

Following the partition standard and considering common relationships in movie scenarios, we annotated five kinds of relationships: workmate relationship, kinship, hostile, friend, and couple relationship. In relationships mentioned above, the workmate relationship corresponds to the hierarchical power domain, while the kinship corresponds to the attachment domain, and the couple relationship corresponds to the mating domain. Particularly, considering that movie plots are usually more refined and dramatic, we add the hostile relationship, while the coalitional groups

domain and the reciprocity domain are combined as the friend relationship to further fit the relationship between characters in movies.

For annotations, we invite six experts who majored in society and human behavior to give relationship annotations between selected character pairs. Watching the whole movie is required before all annotations. Along this line, we determine the relationship annotation between character pairs as the label supported by the most experts. To further improve the reliability of the relationship annotations, the relationship annotations that are supported by less than three experts have been removed from the dataset.

REFERENCES

[1] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multim. Syst.* 16, 6 (2010), 345–379.

[2] Slawomir Bak and Peter Carr. 2016. Person re-identification using deformable patch metric learning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[3] Daphne Blunt Bugental. 2000. Acquisition of the algorithms of social life: A domain-based approach. *Psychol. Bull.* 126, 2 (2000), 187.

[4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6154–6162.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2017. VGGFace2: A dataset for recognising faces across pose and age. https://ieeexplore.ieee.org/abstract/document/8373813.

[6] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2109–2118.

[7] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream CNN model. https://openaccess.thecvf.com/content_ECCV_2018/html/Di_Chen_Person_Search_via_ECCV_2018_paper.html.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1335–1344.

[9] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[11] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recog.* 48, 10 (2015), 2993–3003.

[12] E. Eidinger, R. Enbar, and T. Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forens. Secur.* 9, 12 (2014), 2170–2179. DOI: https://doi.org/10.1109/TIFS.2014.2359646

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.478.6547&rep=rep1&type=pdf.

[14] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *Proceedings of the 26th ACM International Conference on Multimedia*. 1865–1873.

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 87–102.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Qingqiu Huang, Wentao Liu, and Dahua Lin. 2018. Person search in videos with one portrait through visual and temporal links. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 425–441.

[18] Vijay Kumar, Anoop M. Namboodiri, and C. V. Jawahar. 2014. Face recognition in videos by label propagation. In *Proceedings of the 22nd International Conference on Pattern Recognition*. IEEE, 303–308.

[19] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–42.

[20] Chenchen Li, Jialin Wang, Hongwei Wang, Miao Zhao, Wenjie Li, and Xiaotie Deng. 2019. Visual-texual emotion analysis with deep coupled video and Danmu neural networks. *IEEE Trans. Multim.* 22, 6 (2019), 1634–1646.

[21] Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price. 2015. Inner and inter label propagation: Salient object detection in the wild. *IEEE Trans. Image Process.* 24, 10 (2015), 3176–3186.

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.
[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.
[24] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.
[25] Harry T. Reis, W. Andrew Collins, and Ellen Berscheid. 2000. The relationship context of human behavior and development. *Psychol. Bull.* 126, 6 (2000), 844.
[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.
[27] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. 2018. End-to-end deep Kronecker-product matching for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6886–6895.
[28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[29] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Learning representations for multimodal data with deep belief nets. In *Proceedings of the International Conference on Machine learning Workshop*.
[30] Nitish Srivastava and Russ R. Salakhutdinov. 2012. Multimodal learning with deep Boltzmann machines. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 2222–2230.
[31] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3481–3490.
[32] Subarna Tripathi, Serge Belongie, Youngbae Hwang, and Truong Nguyen. 2016. Detecting temporally consistent objects in videos through object class label propagation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.
[33] Tinne Tuytelaars, Marie-Francine Moens et al. 2011. Naming people in news videos with label propagation. *IEEE Multim.* 3 (2011), 44–55.
[34] Fei Wang and Changshui Zhang. 2007. Label propagation through linear neighborhoods. *IEEE Trans. Knowl. Data Eng.* 20, 1 (2007), 55–67.
[35] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6609–6618.
[36] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3415–3424.
[37] Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8280.
[38] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
[39] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sig. Process. Lett.* 23, 10 (2016), 1499–1503.
[40] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. 2015. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4804–4813.
[41] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. CityPersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3221.
[42] Shanshan Zhang, Jian Yang, and Bernt Schiele. 2018. Occluded pedestrian detection through guided attention in CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6995–7003.
[43] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. 2015. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 4678–4686.
[44] Zhedong Zheng, Liang Zheng, and Yi Yang. 2018. Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. Circ. Syst. Vid. Technol.ogy* 29, 10 (2018), 3037–3045.
[45] Olga Zoidi, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. 2014. Person identity label propagation in stereo videos. *IEEE Trans. Multim.* 16, 5 (2014), 1358–1368.

Received November 2020; revised June 2021; accepted August 2021